Designing Representation Learning Applications for Unlabelled Image Data using Visual Foundational Models

Inaugural thesis

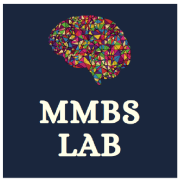
presented to the Faculty of Mathematics and Natural Sciences
of Heinrich Heine University Düsseldorf
for the degree of
Doctor of Philosophy in Natural Sciences

by

Nikolaos Adaloglou
from Veroia, Greece

Düsseldorf, Germany

January 27, 2026



from the Institute of Mathematical Modeling of Biological Systems (MMBS)
of Heinrich Heine University Düsseldorf

Printed by permission of the Faculty of
Mathematics and Natural Sciences of
Heinrich Heine University Düsseldorf

Examiners:
1. *Prof. Dr. Markus Kollmann* (supervisor)
2. *Prof. Dr. Timo Dickscheid* (mentor)

Date of the oral defense: *19th of January 2026.*

Statement of authorship

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the "Principles for the Safeguarding of Good Scientific Practice" at Heinrich Heine University Düsseldorf. No other person's work has been used without due acknowledgment. This dissertation has not been submitted in the same or similar form to other institutions. Results from co-authored publications have been used with proper acknowledgment of collaborators. I have not previously failed a doctoral examination procedure.

Düsseldorf, Germany, 2026
Nikolaos Adaloglou

Acknowledgements

So little time, so much to learn. I have spent four and a half productive years in the Institute of Mathematical Modeling of Biological Systems. During that time, my collaborators and I have authored eight research papers in respectable conferences and journals. The confidence to pursue all these research projects is mainly attributed to the trust and confidence my supervisor had in me. I am eternally grateful to have had *Prof. Markus Kollmann* as my supervisor, because he believed in my skills from day one. Thank you, Markus!

My family has been extremely supportive of me during my PhD candidacy. Thank you for being so kind to me, *Sofia, Maria, Tonia, Artemis, and of course, my mother, Chrisoula.* In memory of my father, Michalis, and grandparents, Maria and Savvas, I wish to express my gratitude to my family by dedicating this thesis to them.

My colleagues at the University of Düsseldorf played a pivotal role in helping me grasp challenging concepts through intense discussions and constructive disagreements. Together, we have developed teaching material for MSc students and even developed state-of-the-art methods without prior knowledge of the particular task. This zero-to-one venture lies at the core of my learning experience. Thank you, *Felix Michels, Kaspar Senft, Asker Mohamed Ahmed, Diana Pertusheva, Damir Iagudin, and Tim Kaiser.*

Good dear friends who have inspired me to do my very best include, but are not limited to, *Ilias Iliopoulos, Katerina Tzovla, Segios Karagiannakos, Manolis Koumantakis, George Zoumpourlis, and Antonis Taliadouros.* Life is so much easier with you! I hope our relationship ages like a good bottle of red wine.

To my future self, I want to let you know I feel proud of the final result. May you be reaping the rewards of this journey, and keep trusting your curiosity and intuition. While I have managed to bring most of my current best ideas to the world, I remain optimistic that the best is yet to come!

Abstract

Visual representation learning lies at the core of foundational AI models for computer vision. Current approaches typically involve learning general-purpose representations from large-scale datasets and adapting them to downstream tasks such as image segmentation, generative modeling, and object detection. However, adapting these models to domains with limited or no labeled data remains a major challenge. This manuscript addresses this gap by developing methods that repurpose visual and multi-modal foundational models for tasks where labels are scarce, including image clustering, visual out-of-distribution (OOD) detection, and label-free generative learning.

Detecting anomalous or out-of-distribution samples is essential for ensuring reliability in safety-critical applications. Similarly, image clustering is a longstanding task that, in addition to identifying useful structures in unlabeled data, can serve as an annotation proxy (pseudo-labels). Pseudo-labels can be leveraged in a plethora of follow-up tasks, such as semi-supervised learning. Finally, visual generative learning without labels, known as unconditional image synthesis, can offer useful solutions in fields such as medical imaging, where expert-level annotations are scarce, expensive, and time-consuming.

Towards repurposing foundational models for downstream tasks, this work makes four main contributions. First, it introduces a novel image clustering approach based on pointwise mutual information and presents a large-scale empirical study of clustering methods under challenging data distributions. Second, it proposes multiple strategies for visual OOD detection, utilizing the availability of ground truth labels, label names, or a sufficiently large text corpus with vision-language models. Third, it demonstrates how representation learning can enhance label-free generative modeling by using cluster assignments to improve training efficiency and inference-time control. Finally, it presents a novel sampling method for both conditional and unconditional diffusion models. The introduced method enhances long-range dependencies during sampling, resulting in visually coherent samples with high perceptual quality.

Collectively, the above contributions advance the adaptation of foundational models to visual label-deprived tasks, providing practical solutions for designing robust, data-efficient computer vision systems.

Zusammenfassung

Das visuelle Repräsentationslernen bildet den Kern von KI-Foundation-Modellen für Computer Vision. Aktuelle Ansätze umfassen in der Regel das Lernen allgemeiner Repräsentationen aus großen Datensätzen und deren Anpassung an spezifischere Aufgaben wie Bildsegmentierung, generative Modellierung und Objekterkennung. Die Anpassung dieser Modelle an Bereiche mit begrenzten oder gar keinen annotierten Daten bleibt jedoch eine große Herausforderung. Diese Arbeit befasst sich mit dieser Lücke, indem sie Methoden entwickelt, die visuelle und multimodale Foundation-Modelle für Aufgaben wiederverwenden, bei denen nur wenige Label vorhanden sind, darunter Bild-Clustering, visuelle Out-Of-Distribution (OOD) Erkennung und generatives Lernen ohne Labels.

Die Erkennung anomaler oder OOD Stichproben ist für die Gewährleistung der Zuverlässigkeit in sicherheitskritischen Anwendungen von entscheidender Bedeutung. In ähnlicher Weise ist das Bild-Clustering eine wichtige Aufgabe, die neben der Identifizierung nützlicher Strukturen in nicht annotierten Daten auch als Annotationsproxy (Pseudo Labels) dienen kann. Pseudo Labels können in einer Vielzahl von Folgeaufgaben (downstream tasks), wie zum Beispiel dem halbüberwachten (semi-supervised) Lernen, genutzt werden. Schließlich kann visuelles generatives Lernen ohne Labels (unconditional image synthesis) nützliche Lösungen in Bereichen wie der medizinischen Bildgebung bieten, in denen Annotationen auf Expertenniveau selten, teuer und zeitaufwendig sind.

Im Hinblick auf die Umnutzung von Foundation-Modellen für Folgeaufgaben leistet diese Arbeit vier wesentliche Beiträge. Erstens wird ein neuartiger Ansatz zur Bildclusterung auf der Grundlage punktweiser gegenseitiger Information vorgestellt und eine groß angelegte empirische Studie zu Clustering-Methoden mit herausfordernden Datenverteilungen präsentiert. Zweitens werden mehrere Strategien für die visuelle OOD-Erkennung entwickelt, wobei die Verfügbarkeit von Ground-Truth-Labels, Label-Namen oder einem ausreichend großen Textkorpus mit Bild-Sprach-Modellen genutzt wird. Drittens wird gezeigt, wie das Repräsentationslernen die labelfreie generative Modellierung verbessern kann, indem Clusterzuweisungen zur Verbesserung der Trainingseffizienz und der Steuerung während der Inferenz verwendet werden. Schließlich wird eine neuartige Sampling-Methode für Diffusionsmodelle mit und ohne Konditionen vorgestellt. Die vorgestellte Methode verbessert die globalen Abhängigkeiten während des Samplings und erzeugt visuell kohärente Samples mit hoher wahrgenommene Qualität.

Insgesamt fördern die oben genannten Beiträge die Anpassung von Foundation-Modellen an visuelle Aufgaben ohne Labels und bieten praktische Lösungen für die Entwicklung robuster, dateneffizienter Computer-Vision-Systeme.

Milestones

By the end of his PhD candidacy, the author achieved the following milestones:

- The author has mentored eight MSc students from the AI & Data science master's program at the Heinrich Heine University of Düsseldorf for their theses: *Diana Petrusheva, Asker Mohamed Raafat Elsayed Ahmed, Hamza Kalisch, Leon Erbrich, Fabian Becker, Malek al-Dineh, Leon Markwart, and Kaspar Senft.*
- To date, the author has written 10 papers and more than 30 blog posts since 2020, resulting in more than 620 citations. Prestigious conferences and journals include *ECCV 2020, TIMM 2021, BMVC 2023, TMLR 2024, WAVC 2025, BMVC 2025 (oral), WACV2026.*
- Together with *Felix Michels*, the author has designed and co-taught the MSc course on representation learning. Lectures and course material are publicly available.
- Together with *Felix Michels* and *Tim Kaiser*, the author created teaching materials and tutored in generative modeling (2022, 2025) and deep learning (2022, 2023).
- Community service: the author has served as a reviewer for a number of prestigious computer vision conferences and journals, including CVPR, ICCV, BMVC, WAVC, and TMLR.
- In parallel with his candidacy, the author successfully completed an 8-month internship in the Machine Learning research department of Bayer AG in 2022.
- The paper "Guiding a diffusion model with itself using sliding windows", accepted in BMVC 2025 (oral), received the **best paper award**.

Keywords

clustering, image clustering, out-of-distribution detection, diffusion models, label mining, image generation, image synthesis, anomaly detection, diffusion guidance, generative conditioning, representation learning

Contents

Part I: Research Synthesis & Overview

Representation Learning **10**
1.1 Notation and Prerequisites . . . 10
1.2 Introduction and Supervised Learning . . . 11
1.3 Mask Modeling: Language and Vision . . . 14
1.4 Contrastive Learning . . . 16
1.5 Joint-Embedding Self-Supervised Approaches . . . 23
1.6 Image Representations *Without* Hand-crafted Augmentations . . . 30
1.7 Representation Convergence . . . 31
1.8 Generative Learning using Visual Representations . . . 33
Downstream Tasks that Leverage Visual Representations **35**
2.1 Visual Out-of-Distribution Detection . . . 35
2.2 Large-Scale Image Clustering . . . 45
2.3 Image Synthesis with Diffusion Models . . . 51
Contributions and References **63**
Contributions . . . 63
References . . . 64

Part II: Core Publications

Exploring the Limits of Deep Image Clustering **94**
A Introduction . . . 1
B Related Work . . . 2
C Proposed Method . . . 3
D Experimental Evaluation . . . 7
E Conclusion . . . 11
References . . . 13
Appendix . . . 17
Adapting CLIP for Out-of-Distribution Detection **21**
A Introduction . . . 1
B Related Work . . . 3
C Proposed OOD Detection Setup . . . 4

D Experimental Results . . . 10
E Discussion . . . 13
F Conclusion . . . 14
References . . . 16
Appendix . . . 24
Rethinking Cluster-Conditioned Diffusion Models **27**
A Introduction . . . 1
B Related Work . . . 2
C Method . . . 3
D Experimental Evaluation . . . 6
E Discussion . . . 10
F Conclusion . . . 15
References . . . 16
Appendix . . . 22
Scaling Up Feature-Based Image Clustering Methods **26**
A Introduction . . . 1
B Related Work . . . 2
C Background, Materials and Methods . . . 3
D New Clustering Benchmarks . . . 5
E Experimental Results . . . 7
F Discussion and Limitations . . . 10
G Conclusion . . . 12
References . . . 13
Appendix . . . 18
ClusterMine: Robust Label-Free Visual Out-Of-Distribution Detection via Concept Mining from Text Corpora **25**
A Introduction . . . 1
B Related work . . . 2
C OOD detection methods using CLIP . . . 3
D Experimental evaluation . . . 6
E Future work and conclusion . . . 13
References . . . 15
Guiding a Diffusion Model with Itself using Sliding Windows **21**
A Introduction . . . 1
B Related work . . . 2
C CFG and WMG are fundamentally different . . . 3
D Sliding window guidance (SWG) . . . 5
E Experimental evaluation and discussion . . . 7
F Conclusion . . . 11
References . . . 12
Appendix . . . 15
Conclusion and Publication list **17**

Part I:
Research Synthesis & Overview

1. Representation Learning

1.1 Notation and Prerequisites

Let X be a discrete random variable with sample space $\mathcal{X}$ with probability mass function $p(x) = Pr\{X = x\}, x \in \mathcal{X}$, where x is the observed value, also known as a realization of $X \sim p(x)$ [56]. We denote the probability mass function $p(x)$ of the discrete random variable X, where $\sum_{x\in\mathcal{X}} p(x) = 1$. The sample space $\mathcal{X}$ is the set of all possible values, and it is also known as the alphabet or support in information theory [182], or the state space in system modeling. Similarly, two random variables X and Y with a joint probability distribution $p(x, y)$ and marginal probability distributions $p(x)$ and $p(y)$ can be defined. Observations $x, y \in \mathcal{X} \times \mathcal{Y}$ can be sampled from the joint distribution $p(x, y)$ where $(X, Y) \sim p(x, y)$, or from the marginal distributions $p(x)$ and $p(y)$.

Encoded representations of the observed samples $z = g_\theta(x)$ can also be regarded as realizations of random variables, $Z \sim p(z)$, where $g_\theta(.)$ is a parametrized function with parameters θ (e.g. a deep neural network). Finally, the mutual information $I(X; Y)$ between two discrete random variables X and Y quantifies the amount of information one variable provides about the other

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1.1}$$

$$= \mathbb{E}_{p(x,y)}\left[\log\frac{p(x,y)}{p(x)p(y)}\right] \tag{1.2}$$

$$= H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{1.3}$$

In Eq. (1.3), the *entropy* H(X) of a discrete random variable X with probability mass function $p(x)$ is defined as $H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$, which quantifies the average uncertainty of X. $H(X)$ measures the expected amount of information needed to describe X, across all possible $x \in \mathcal{X}$. The *conditional entropy* of X given Y is defined similarly

$$H(X|Y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y). \tag{1.4}$$

$H(X|Y)$ measures the average uncertainty remaining about X when Y is known. Consequently, $I(X; Y)$ quantifies the *reduction in the uncertainty* of X due to the knowledge of Y and vice versa. The ratio in the logarithm of $I(X; Y)$

$$\frac{p(x,y)}{p(x)p(y)} = \frac{p(x|y)}{p(x)} = \frac{p(y|x)}{p(y)}, \text{ as } p(x,y) = p(x)p(y|x) \tag{1.5}$$

can be understood as a scaled conditional probability. The mutual information can also be viewed as the *relative entropy* between the joint distribution $p(x, y)$ and the product of the marginals

$$I(X;Y) = D_{KL}\left(p(x,y) \parallel p(x)p(y)\right) \geq 0. \tag{1.6}$$

Relative entropy was first defined by Kullback and Leibler (KL), and it is known under a plethora of names. In general, for any probability mass function or probability distribution p, q we will refer to $D_{KL}\left(p\|q\right)$ (Eq. (1.6)) as *KL divergence* in this manuscript. Due to its asymmetric nature, $D_{KL}\left(p\|q\right) \neq D_{KL}\left(q\|p\right)$. KL divergence is *a measure of inefficiency* of estimating the true distribution p (the target), given that the estimated distribution is q (the prediction). For instance, the distribution q can be computed via a neural network that optimizes the parameters θ with respect to an objective function via backpropagation.

1.2 Introduction and Supervised Learning

The performance of learning algorithms heavily depends on the nature of their input representations [38]. Representation learning methods seek a representation of the joint distribution over events in the world that generate the data we observe [123]. The observed data are measured by sensors (e.g. a camera) or other projections of the real world, such as textual descriptions (natural language). Representation learning algorithms transform the raw observation $x \in \mathcal{X}$ into a typically lower-dimensional representation z that statistically models x. A good representation retains relevant information for one or more downstream tasks [157, 84].

Outside of machine learning, a classic example can be found in data compression, where only the most informative aspects of raw sensory data are preserved. Most bits of information in perceptual signals (e.g. pixels) carry minimal perceptual relevance [34, 23], while our audiovisual senses are more sensitive to low frequencies [68]. Lossy compression algorithms, such as JPEG and MP3 [274, 127], exploit redundancies in perceptual signals, enabling us to represent them with fewer bits. Instead of relying on handcrafted algorithms, representation learning uses observations to learn z. Large-scale language modeling has been associated with compression [62, 185].

Supervised representation learning. The first major advancement in representation learning was achieved through supervised learning. In this setup, the labels y are known while $p(y|x)$ is learned [152, 111]. Typically, a neural network with learnable parameters θ is trained to approximate $p(y|x)$. At test time, an intermediate representation z of the neural network is used, such as the penultimate layer. This paradigm is commonly referred to as supervised representation learning. The representation z can be used as is (i.e., train a task-specific neural network using z as input) or the weights θ can be refined through fine-tuning, also known as transfer learning [321].

Transfer learning enables knowledge transfer from a source distribution with ample labeled samples to a target distribution [83]. When the source and target distributions are sufficiently related, transfer learning can enhance target task performance and improve data efficiency [321]. For example, the ImageNet accuracy is highly correlated with other

image recognition tasks, though the correlation weakens for object detection [75]. When the number of labeled samples is low, the representations learnt by supervised methods exhibit poor generalization to samples outside the training distribution [145]. In such scenarios, additional objectives, such as autoencoding, can be employed as an auxiliary regularization to increase data efficiency [147, 197].

Collecting and annotating sufficient training data for supervised learning is frequently expensive, error-prone, and time-consuming [321, 30]. The infamous ImageNet21K dataset [63], a considerable effort in collecting annotated image samples, required over 25,000 workers to annotate 14 million images for 22,000 semantic categories [214]. Moreover, human biases in the annotation procedure can lead to systematic inaccuracies [4, 30]. To address these limitations, weaker supervision strategies that reduce reliance on manual human annotation have been proposed. While inherently noisy, weaker forms of supervision, such as web-scraped image descriptions or hashtags, can provide a useful supervisory learning signal when obtained at scale [184, 214].

One of the primary challenges in leveraging noisy paired data for representation learning lies in effective data curation [81, 82, 273] or improving the associated labels per sample [29]. Another approach [242, 241, 184] involves obtaining hashtags from associated free-form image captions and mapping them to WordNet synonym sets (SynSets) [189]. WordNet is a lexical database that organizes words into sets of synonyms (SynSet) and arranges them in a hierarchical, tree-like structure based on semantic relationships [189]. A SynSet represents a group of cognitive synonyms that convey a shared concept or meaning. In the current largest-scale supervised pre-training on image data [60], 4B images have been semi-automatically annotated to a class hierarchy of 30k labels for multi-label classification. Dehghani et al. [60] report an unprecedented robustness and alignment to human perception [90].

Connection to mutual information. Supervised learning can be viewed under the prism of the information bottleneck principle [262]. By discarding superfluous features with respect to the task (e.g. noise, spurious correlations), the learning algorithm is forced to encode the essential and invariant features. Recent studies reveal a correlation between the degree of information bottleneck and the generalization error [139]. The information bottleneck principle aims to minimize the mutual information between X, Y through observations $x \in \mathcal{X}$ and their intermediate representation $z \in \mathcal{Z}$, while maximizing the task-relevant information, as measured by the associated labels $Y \sim p(y)$, between Z, Y [27]. Mathematically,

$$\max \left[I(Z;Y) - \beta I(X;Z) \right], \tag{1.7}$$

where $\beta \geq 0$ is the Lagrange multiplier. Large values for β result in a highly compressed representation of $X \sim p(x)$. Essentially, Eq. (1.7) forces Z to act as a minimal sufficient statistic of X for predicting Y, known as the *InfoMin* principle [262]. Without label information about the downstream task, minimizing $I(X, Z)$ means discarding information, which is a more challenging task. Based on Eq. (1.7), supervised learning is not incentivized to learn features beyond the bare minimum necessary to classify a sample [285, 5].

1.2.1 Connection to Autoencoders

Alemi et al. [8] relate the formulation of Eq. (1.7) to the variational auto-encoding (VAE) objective [144, 38]. The latter proof can be found in Appendix B in [8], and can be formulated as:

$$\min_{\theta,\psi} \mathcal{L}_{\text{VAE}}(x) = \min_{\theta,\psi} \Bigg[\underbrace{-\mathbb{E}_{q_\psi(z|x)} \left[\log p_\theta(x|z)\right]}_{\text{Reconstruction error}} + \beta \underbrace{D_{\text{KL}} \left(q_\psi(z|x) \parallel p(z)\right)}_{\text{Regularization}}, \Bigg] \tag{1.8}$$

where $q_\psi(z|x)$ and $p_\theta(x|z)$ are the encoder and decoder of VAE, respectively. The prior distribution $p(z)$ is usually a unit Gaussian [145]. This is equivalent to maximizing:

$$\max_{\theta,\psi} -\mathbb{E}_{p(x)} \mathcal{L}_{\text{VAE}}(x) = \max_{\theta,\psi} \left[\underbrace{\mathbb{E}_{p(x)} \mathbb{E}_{q_\psi(z|x)} \left[\log p_\theta(x|z)\right]}_{\leq I(X;Z)} - \beta \; \mathbb{E}_{p(x)} \mathbb{E}_{q_\psi(z|x)} \left[\log \frac{q_\psi(z|x)}{p(z)}\right] \right]. \tag{1.9}$$

Maximizing $I(X;Z)$ is equivalent to minimizing the conditional entropy, since $I(X;Z) = H(X) - H(X \mid Z)$. In the context of autoencoders, $H(X \mid Z)$ is related to the expected negative log-likelihood [272]. The second regularization term in Eq. (1.9) constrains the capacity of z, which encourages the model to learn an efficient representation. The output of the encoder z is often referred to as the latent representation. In practice, we use samples from the minibatch B and sample a single latent per sample in each iteration (Monte Carlo estimation). The term $D_{\text{KL}} \left(q_\psi(z|x) \parallel p(z)\right)$ can be computed analytically when both distributions are Gaussians.

The hyperparameter β determines the tradeoff between image reconstruction quality and latent space compactness [265]. The compactness of this refers to how tightly and efficiently it encodes information [68]. Below are use-cases for β:

- **ELBO:** For $\beta = 1$, the VAE objective is bounded by the evidence lower bound (ELBO), which maximizes the likelihood of the data [144].
- **Interpretability ($\beta > 1$):** When optimizing for compactness and semantic interpretability of the representation, increasing $\beta > 1$ is desirable as in β-VAEs [38]. Increasing β generally comes at the cost of inferior reconstruction quality.
- **Unregularized autoencoder ($\beta = 0$):** the reconstruction training objective of unregularized autoencoders has been linked to the maximization of a lower bound of $I(X, Z)$ [272]. Linsker et al. [171] first described this as *InfoMax*, or the principle of maximum information preservation. This is in contrast to supervised learning, as shown in Eq. (1.7). When $\beta = 0$, architectural constraints are imposed to avoid the trivial mapping of $x = z, \forall x \in \mathcal{X}$ [272]. The per-pixel reconstruction error (mean square) assumes pixel independence, and it usually leads to blurry reconstructions [48]. The model focuses on accurately predicting low frequencies compared to high ones (e.g., texture), as low frequencies constitute the largest fraction of the total signal power [67].

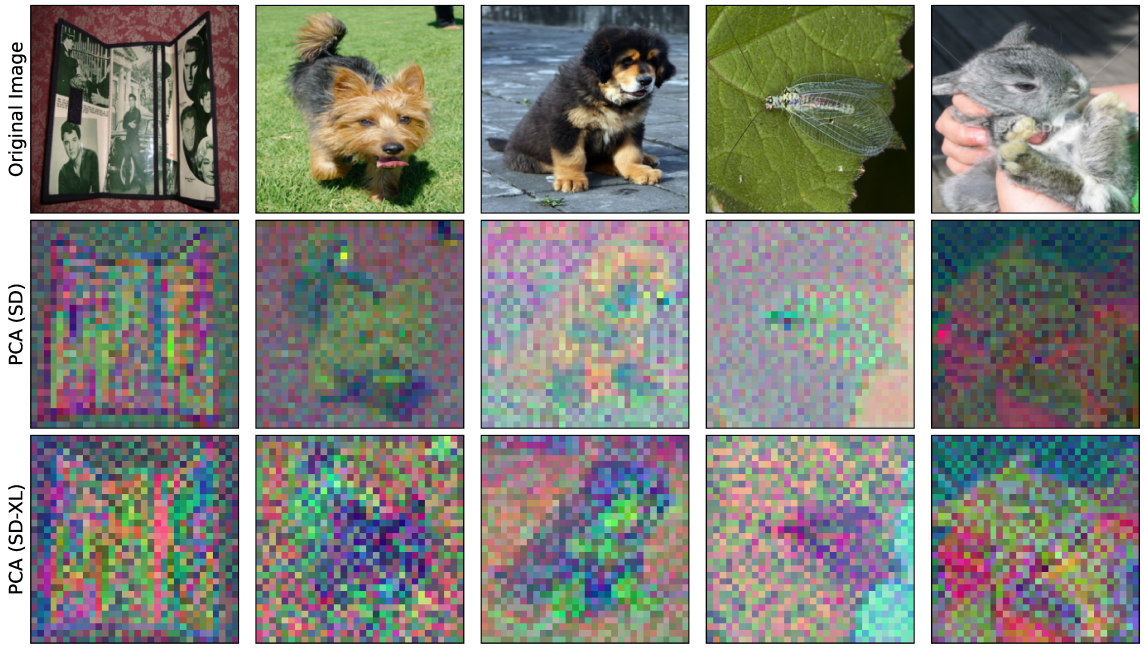


Figure 1.1: **PCA 3-channel visualization of the latent representations as encoded from the VAE [224].** ImageNet samples resized to 256×256 resolution are used without noise, and the VAE weights are based on the stable diffusion (SD) weights from [210], available on the Hugging Face Diffusers library, which encode the image in four channels with a spatial compression factor of 8. SD-XL refers to the largest scale model.

- **Structure preservation ($\beta \in (0, 1)$):** By contrast, for state-of-the-art visual generation using latents [224], VAEs are typically trained with $\beta < 1$ to partially preserve the structure of the image. We visualize this in Fig. 1.1. The retained latent structure can then be exploited with an additional generative model with the appropriate inductive biases [68].

Moreover, it is known that the reconstruction error alone is insufficient for learning useful representations [27]. The autoencoder may memorize fine-grained details irrelevant to the data-generating process. To this end, several regularization techniques have been introduced [271, 145, 239, 149]. For instance, current variational autoencoders (Fig. 1.1) employ a perceptual loss [314], a patch-based adversarial loss [126], and KL regularization with $\beta < 1$ [144]. The above example highlights the difficulty of designing *unified representations* that can be used for discriminative and generative visual tasks. Surprisingly, denoising autoencoders [271] have recently gained traction towards creating such representations [290, 52].

1.3 Mask Modeling: Language and Vision

BERT. A key approach to learning general-purpose representations without labels is masking. As the name suggests, in mask modeling, parts of the input are masked. During training, the model tries to predict the masked segments, such as missing words in a sentence or patches in an image. Mask-based modeling is a widely established objective

in natural language processing (NLP) with remarkable success [65, 216, 36]. In NLP, masking can be applied in random segments of the input sentence, allowing the model to leverage *bidirectional* context to predict the masked segment, known as BERT [65]. In BERT [65], the two views of a sentence correspond to a masked word in the sentence and its bidirectional context. Kong et al. [148] link the BERT objective to a lower bound on the mutual information between different parts of a sentence.

GPT. Another popular approach is left-context-only masking, which explicitly predicts the subsequent parts of a sentence [36], called generative pretraining (GPT). GPT can be efficiently implemented in the transformer architecture using constrained self-attention. The latter is referred to as a "transformer decoder" and can be used autoregressively at inference time. Generative pretraining has partially bridged the gap between generation and recognition in NLP [215, 216].

Masking in vision. However, masked *image* modeling (MIM) methods usually achieve competitive results in discriminative tasks *only through fine-tuning* [100, 317, 205, 241, 160]. For instance, reconstructing a masked patch does not necessarily teach the model to recognize objects [117]. MIM encourages models to focus on local patterns across all layers and is better suited for geometric, motion, and fine-grained tasks [292]. Exemplar tasks include pose estimation, image segmentation, depth estimation, and video object tracking. Park et al. [205] argue that self-supervised vision transformers trained with MIM representations capture more high-frequency information (primarily in early layers), compared to contrastive learning. The most widely used MIM approaches are: a) masked autoencoders (MAE) [144, 100], b) BERT-like image pre-training [20, 208], and more recently, denoising (diffusion) autoencoders [281, 290, 52]. Conceptually, denoising noised image patches is analogous to predicting masked patches in MAE [281].

Natural language versus vision and unified learning. In general, the observed discrepancy in language compared to vision is attributed to multiple factors:

- Information density varies. Languages are human-generated signals that are highly semantic and information-dense [100]. Predicting missing words in a sentence induces linguistic understanding. Conversely, images are natural signals with substantial spatial redundancy. For this reason, masking image modeling often requires higher masking ratios than natural language [65, 284].

- The continuous nature of pixel intensities makes it harder to define a proper "vocabulary" for prediction, unlike the discrete nature of languages.

- Image data has strong locality as nearby pixels are highly correlated. Masking parts of an image can result in learning local details rather than global image understanding.

Recent methods like *data2vec* [17] have unified learning across modalities by only employing modality-specific *tokenization*. Tokenization [269] refers to the process of breaking down raw input data into smaller, discrete units called tokens, which are then mapped to numerical representations. The nature of these tokens depends on the input modality. For example, text may be divided into words or characters; image tokens might correspond to patches.

In this work, we will mainly focus on *discriminative representations* from images and multi-modal data that do *not* require fine-tuning. Such off-the-shelf representations

have been adapted to downstream tasks, such as image clustering and out-of-distribution detection [6]. This approach is significantly more computationally efficient. For each downstream task, it allows training a relatively small model on the learned representations [4] or only partial fine-tuning of the foundational model [5].

1.4 Contrastive Learning

Another principled approach for representation learning, grounded in information theory, is contrastive learning. It encourages the retention of semantically consistent features [84]. Formally, contrastive learning aims to learn a representation of a real concept based only on co-occurring observations [124]. In the following, we consider two co-occurring samples $(v_1, v_2) \in \mathcal{V}_1 \times \mathcal{V}_2$, drawn from the joint distribution $(V_1, V_2) \sim p(v_1, v_2)$ as *positives*. For instance, (v_1, v_2) can be coinciding stochastic views of the same underlying sample x, where x is a realization of $X \sim p(x)$. Each view is obtained by an independent stochastic transformation of x, denoted as $v = \mathcal{T}(x)$, so that

$$p(v_1, v_2 \mid x) = p(v_1|x)\; p(v_2|x). \tag{1.10}$$

Taking the expected value with respect to X then yields the joint distribution of co-occurring views:

$$p(v_1, v_2) = \sum_{x \in \mathcal{X}} \underbrace{p(v_1 \mid x)\; p(v_2 \mid x)}_{p(v_1, v_2|x)}\; p(x). \tag{1.11}$$

For the multi-modal setting, we illustrate the formulation using two paired modalities, such as image-text pairs. In general, contrastive learning can be extended to more modalities by anchoring pairs on one reference modality [259]. In the two-modality case, let (x_1, x_2) denote paired instances of $(X_1, X_2) \sim p(x_1, x_2)$. Each instance is then mapped to a stochastic view via modality-specific independent transformations $\mathcal{T}_1, \mathcal{T}_2$, yielding a positive pair

$$v_1 = \mathcal{T}_1(x_1), \quad v_2 = \mathcal{T}_2(x_2). \tag{1.12}$$

Absence of a transformation corresponds to the identity mapping, i.e. $\mathcal{T}(x) = x$. Since the transformations act independently, we have

$$p(v_1, v_2 \mid x_1, x_2) = p(v_1 \mid x_1)\, p(v_2 \mid x_2). \tag{1.13}$$

Taking an expectation with respect to paired data (x_1, x_2) then gives the joint distribution of positive multi-modal samples:

$$p(v_1, v_2) = \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \underbrace{p(v_1 \mid x_1)\, p(v_2 \mid x_2)}_{p(v_1, v_2|x_1, x_2)}\; p(x_1, x_2). \tag{1.14}$$

Multi-modal realizations include 2D projections of the same 3D scene or observations from different visual sensors [260]. To construct *negative* pairs, we consider independently sampled views $\tilde{v}_1, \tilde{v}_2$ drawn from the product of marginals:

$$p_{\text{neg}}(\tilde{v}_1, \tilde{v}_2) = p(\tilde{v}_1)\; p(\tilde{v}_2), \qquad \text{where} \qquad p(v) = \sum_{x \in X} p(v \mid x)\, p(x), \tag{1.15}$$

which represents unrelated negative pairs. To distinguish samples from the joint distribution $p(v_1, v_2)$ from these negative pairs, a "critic" is defined. The critic refers to a discriminating function $h_\theta(.)$ parametrized by two neural networks with parameters $\theta = \{\theta_1, \theta_2\}$, which may or may not be shared. The respective encoders $g_{\theta_1}(.), g_{\theta_2}(.)$ map views to latent representations

$$z_1 = g_{\theta_1}(v_1), \quad z_2 = g_{\theta_2}(v_2), \qquad z_1, z_2 \in \mathbb{R}^d. \tag{1.16}$$

Subsequently, normalization is applied, positing vectors in the unit hyper-sphere ($z_i \leftarrow \frac{z_i}{\|z_i\|}$), meaning they can only differ by pointing in different directions. The critic is typically implemented as an exponentiated, temperature-scaled cosine similarity

$$h_\theta(v_1, v_2) = \exp\left(\tfrac{1}{\tau} \cdot \tfrac{z_1^\top z_2}{\|z_1\|\|z_2\|}\right) = \exp\left(\frac{1}{\tau}\,\mathrm{sim}(\mathrm{z}_1, \mathrm{z}_2)\right) > 0, \qquad \tau > 0. \tag{1.17}$$

The encoder parameters θ_2 are often identical to θ_1 (siamese networks [50, 51]), or updated as an exponential moving average (EMA) of θ_1 [101]. The EMA requires the two encoders to have an identical architecture, and the parameter update is applied element-wise:

$$\theta_2 \leftarrow \gamma\,\theta_2 + (1-\gamma)\,\theta_1, \qquad \gamma \in (0, 1), \tag{1.18}$$

where values of γ close to 1 are often used. For a given anchor view v_1 (as a reference), we consider a candidate set denoted as

$$s_{v_1} = \{v_1, v_2^1, v_2^2, \ldots, v_2^{k+1}\}, \tag{1.19}$$

where v_2^1 denotes the positive view associated with the anchor view v_1, while $k = |s_{v_1}| - 2$ denotes the number of negative pairs. Negatives are drawn from the marginal distribution $p(v)$, hence

$$p_{\mathrm{neg}}(v_1, v_2^i) = p(v_1)\,p(v_2^i), \text{ for all } i \in \{2, \ldots, k\}, \tag{1.20}$$

similar to Eq. (1.15). The contrastive objective for a single positive pair (v_1, v_2^1), can then be written as the negative log-softmax over all k negative pairs as:

$$\mathcal{L}_{\mathrm{contrast}}(s_{v_1}) = -\log \frac{\exp\left(\mathrm{sim}(z_1, z_2^1)/\tau\right)}{\sum_{j=1}^{k+1} \exp\left(\mathrm{sim}(z_1, z_2^j)/\tau\right)} = -\log \frac{h_\theta(v_1, v_2^1)}{\sum_{j=1}^{k+1} h_\theta(v_1, v_2^j)}. \tag{1.21}$$

By sampling multiple s_{v_1} i.i.d. sets, denoted as $\mathcal{S}_{v_1}$, we can extend the above equation as:

$$\mathcal{L}_{\mathrm{contrast}}(S_{v_1}) = \mathcal{L}_{\mathrm{contrast}}(V_1, V_2^1) = -\frac{1}{|\mathcal{S}_{v_1}|} \sum_{s_{v_1} \in \mathcal{S}_{v_1}} \mathcal{L}_{\mathrm{contrast}}(s_{v_1}). \tag{1.22}$$

Intuitions. An intuitive way to illustrate the contrastive objective is by decomposing and approximating Eq. (1.21) as two terms using a shared encoder g:

$$\mathcal{L}_{\text{contrast}} \approx \underbrace{-\frac{1}{\tau}\frac{g(v_1)^\top g(v_2^1)}{\|g(v_1)\| \cdot \|g(v_2^1)\|}}_{\text{alignment/invariance}} + \underbrace{\log \sum_{j=2}^{k+1} \exp\left(\frac{g(v_1)^\top g(v_2^j)}{\|g(v_1)\| \cdot \|g(v_2^j)\|} \cdot \frac{1}{\tau}\right)}_{\text{uniformity/regularization}}. \tag{1.23}$$

The first term aligns the features from positive pairs in the d-dimensional hypersphere [279]. Another perspective relates the first term to learning invariances between positive pairs. The second term spreads the features from negative pairs roughly uniformly on the hypersphere as $k \to \infty$ [279]. In other words, negative samples are distinguishable in feature space [26]. The regularization term primarily drives the grouping of samples with respect to their semantic labels. In general, current self-supervised methods require a regularization constraint to prevent representation collapse, wherein the model maps all inputs to the same output [132]. The contrastive loss can also be viewed as a $k+1$ softmax classification problem. Similarities between representations act analogously to the output of a supervised classifier, also known as logits [50].

Implementing the contrastive loss. In practice, k can be arbitrarily large. Thus, directly minimizing Eq. 1.21 is infeasible. The most widely adopted solution is to randomly sample $k = 2B - 2$ negatives, given a mini-batch of B samples with two views per sample. Negative pairs are computed from the similarities between non-co-occurring representations, as illustrated with the negative rectangles in Fig. 1.2. The mini-batch similarity matrix $\mathbf{Z}$ is defined as:

$$\mathbf{Z}\mathbf{Z}^\top \in R^{2B\times 2B}, \mathbf{Z} = \left[\frac{z_1}{||z_1||}, \ldots, \frac{z_{2B}}{||z_{2B}||}\right] \in \mathbb{R}^{2B\times d}. \tag{1.24}$$

Consequently, for all $2B$ positive pairs in the mini-batch indexed by i, j, as highlighted with green in Fig. 1.2, the contrastive loss is formulated as:

$$\mathcal{L}_{contrast}(v_i, v_j) = -\log \frac{h_\theta(v_i, v_j)}{\underbrace{\sum_{k=1}^{2B} \mathbb{1}_{[k\neq i]} h_\theta(v_i, v_k)}_{\text{postitive + negatives}}}, \tag{1.25}$$

where $\mathbb{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 0 when $k = i$ to exclude self-similarities (black filled rectangles in Fig. 1.2), which correspond to $\text{diag}(\mathbf{Z}\mathbf{Z}^\top)$. We highlight that the denominator in Eq. (1.25) includes the positive pair (v_i, v_j) in the nominator. The final loss is averaged across all positive pairs in the mini-batch [245, 288, 201]:

$$\mathcal{L}_{contrast} = \frac{1}{2B}\left(\sum_{i=1}^{B} \mathcal{L}_{contrast}(v_{2i-1}, v_{2i}) + \sum_{i=1}^{B} \mathcal{L}_{contrast}(v_{2i}, v_{2i-1})\right). \tag{1.26}$$

The above loss is referred to as normalized temperature-scaled cross-entropy loss (NT-Xent) or InfoNCE. Another well-known alternative is MoCo [101], which keeps a memory

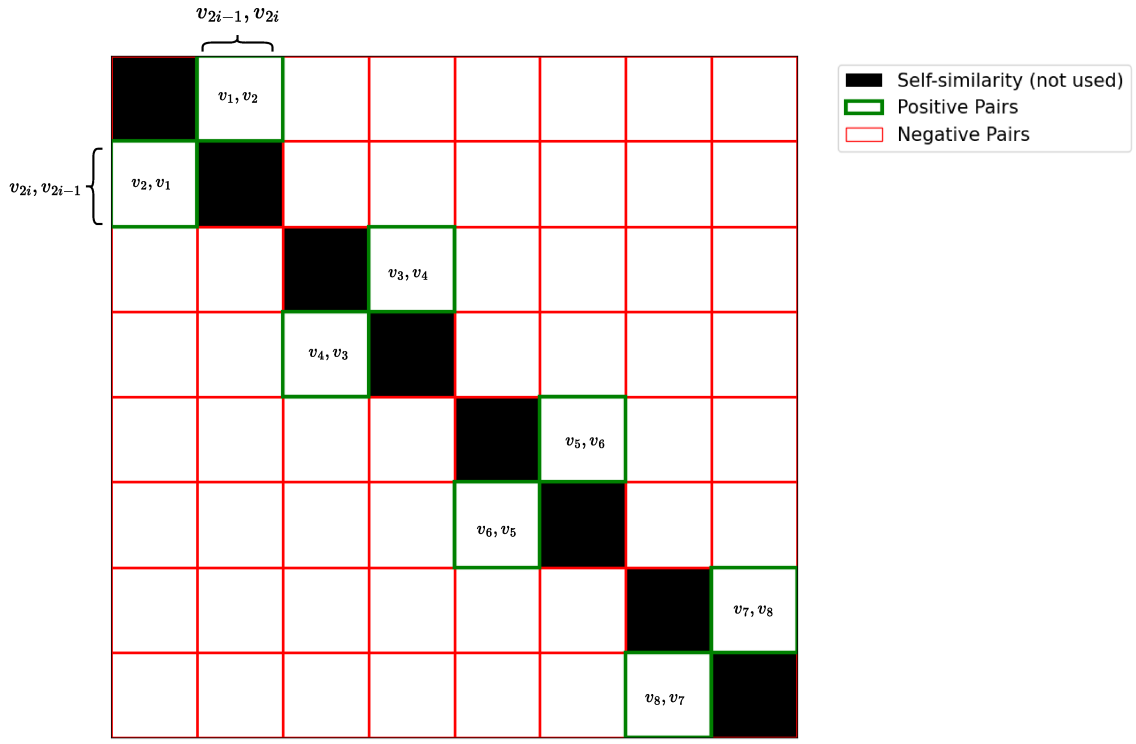


Figure 1.2: **A graphical illustration for $B = 4$ of the positive and negative pairs within a mini-batch, as used in the SimCLR/InfoNCE objective.**

bank of k samples from previous iterations as negatives, combined with an exponential moving average (EMA) of the encoder's parameters (Eq. (1.18)). The memory bank is dynamically updated with features computed on each iteration. With this approach, the negative pairs are independent of the batch size, i.e. $k >> 2B - 1$. After the training phase, one of the encoders, $g_{\theta_1}(.), g_{\theta_2}(.)$, is kept for the downstream task.

Connecting the contrastive objective to mutual information. Contrastive learning can be interpreted as a classification task where, for a fixed anchor representation z_1, one must identify its paired sample z_2 among $k + 1$ candidates. Here, we are using the representations z, which follow the same distributions as the views and are linked via the encoder $z = g_\theta(v)$. Thus, a positive pair drawn from $p(v_2 \mid v_1)$ is shown as $p(z_2 \mid z_1)$ to emphasize that the optimization takes place in representation space. Similarly, the k negatives are drawn independently from the marginal $p(\tilde{z}_2)$ (Eq. (1.15)). In the population limit ($k \to \infty$), with negatives drawn i.i.d. from the marginal $p(\tilde{z})$, minimizing the expected cross-entropy loss over infinitely many pairs is equivalent to matching $p_\theta(z_2 \mid z_1)$ to the true conditional distribution $p(z_2 \mid z_1)$. Based on that, it can be proved that the optimal critic $h^*_\theta(.)$ is proportional to the ratio [260]:

$$h^*_\theta(v_1, v_2) \propto \frac{p(z_1, z_2)}{p(z_1)p(z_2)} \propto \frac{p(z_1|z_2)}{p(z_1)} = \frac{p(z_2|z_1)}{p(z_2)}, \quad \text{where} \quad z = g_\theta(v). \tag{1.27}$$

This ratio is known as pointwise mutual information (PMI), and its expectation, in Eq. 1.21, yields an estimator related to mutual information $I(Z_1; Z_2)$. To show this

link, we first define the contrastive loss, similarly to Eq. (1.22), for a large sample set of positive pairs as:

$$\mathcal{L}_{\text{NCE}} = -\mathbb{E}_{p(v_1,v_2)}\,\mathbb{E}_{p(\tilde{v}_2^j)}\left[\log \frac{h_\theta(v_1,v_2)}{h_\theta(v_1,v_2)+\sum_{j=2}^{k+1} h_\theta(v_1,\tilde{v}_2^j)}\right], \tag{1.28}$$

where $p(\tilde{v}_2^j)$ follows Eq. (1.15) for all $j > 1$. It can be proved that, for sufficiently large k, $\mathcal{L}_{\text{NCE}}$ in Eq. (1.28) yields a lower bound on mutual information. A proof can be found in [259, 201, 211], showing that:

$$I(Z_1; Z_2) \;\geq\; \log(k+1) \;-\; \mathcal{L}_{\text{NCE}}. \tag{1.29}$$

By the data-processing inequality, $I(Z_1; Z_2)$ is also bounded above by the views, where $I(V_1; V_2)$ is constant:

$$I(V_1; V_2) \geq I(Z_1; Z_2) \geq \log(k+1) \;-\; \mathcal{L}_{\text{NCE}}. \tag{1.30}$$

In practice $\mathcal{L}_{\text{NCE}}$ is approximated by the averaged empirical loss $\mathcal{L}_{\text{contrast}}$ defined in Eq. (1.26) using k negatives from the mini-batch, where $k = 2B - 2$. Thus, minimizing $\mathcal{L}_{\text{contrast}}$ at each iteration is expected to maximize a lower bound on the mutual information between co-occurring representations. In practice, finite-sample mini-batch estimates of $\mathcal{L}_{\text{NCE}}$ can introduce variance and bias, so the bound equation 1.29 may remain loose [211, 187].

1.4.1 Designing Positive Views in Contrastive Learning

InfoMin or InfoMax? When dealing with a single modality, such as natural images, creating positive views is a key design choice. This decision dictates the shared information between (v_1, v_2), which alters the joint distribution and ultimately influences the learned representation [159]. The InfoMin principle [262, 8] suggests that a good set of views should capture the minimal sufficient statistics required for a downstream task. The InfoMax principle emphasizes preserving maximal information [171], which is not downstream task-dependent, unlike InfoMin. Tian et al. [261] state that InfoMin is complementary to the InfoMax principle in contrastive learning. In particular, beyond maximizing all the task-relevant information, discarding irrelevant information tends to improve generalization [235, 121, 243]. Under this prism, both principles converge on the choice of positive pairs. Ideally, the only information shared between (v_1, v_1) is relevant for the downstream task.

When the amount of shared information between (v_1, v_2) is small, then the learned representation will implicitly discard more information about the input [261]. To exemplify, let us consider two *extreme cases*:

1. **Ground truth pairs.** When human labels are present, same-class samples serve as positive pairs, known as *supervised contrastive learning* [140], and only class-related information is learned. Khosla et al. [140] show that comparable results to standard supervised learning with cross-entropy can be obtained using the supervised contrastive objective.

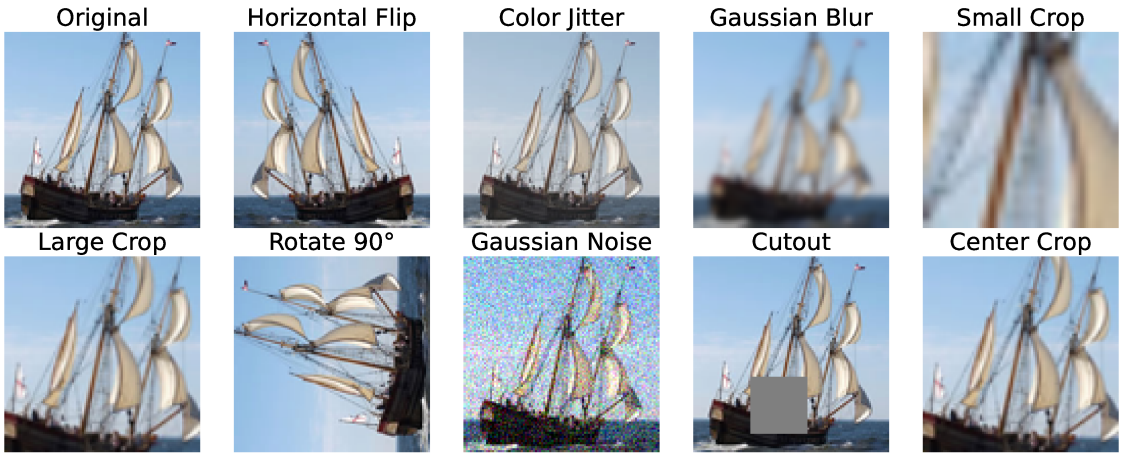


Figure 1.3: **Illustrations of the visual data augmentation operations.** Each augmentation is applied stochastically with hyperparameters such as the crop scale, probability of applying a transformation, etc. STL10 samples are used.

2. **Missing shared information.** While undesirable, views can be designed in such a way to discard critical information, e.g. $I(V_1; V_2) < I(X; Y)$, where X and Y denote the random variables corresponding to the input image and its associated label, respectively. Assuming an object-centric dataset like ImageNet [261], task-relevant information is discarded by considering only small and distant crops that exclude the labeled object.

Image augmentations and learned invariances. When dealing with unlabelled observations from a single modality, the common practice is to use stochastic augmentations. In computer vision, augmentations typically refer to random geometric and photometric transformations, such as horizontal flip and blurring (Fig. 1.3). The first systematic study on image augmentation was conducted by [50]. It was found that random crops and color distortions are the most crucial ones for competitive performance in image recognition tasks. However, reliance on domain-specific augmentations limits the generalizability of representation learning approaches [196, 141]. Still, augmentations are necessary for achieving state-of-the-art performance, especially in low-data regimes.

Nevertheless, augmentations can sometimes push apart semantically consistent samples, which is undesirable for image recognition tasks (Fig. 1.3). This behaviour of contrastive learning is closely tied to the sensitivity of the temperature hyperparameter τ [275]. An excessively low temperature τ disproportionately penalizes similar samples. A good choice of τ balances the alignment of views from the same image while preserving the semantic structure in feature space. Another alternative to address the above limitation is to construct positive views that remain invariant to intra-class variations. Inspired by MoCo [101], NNCLR dynamically retrieves nearest neighbors from a memory bank as likely positive pairs [74].

Representational knowledge distillation. Knowledge distillation can be framed as a contrastive learning problem [48]. In standard knowledge distillation, knowledge is transferred from a trained teacher model to a student model [112], typically by minimizing

the KL divergence between their output distributions. This approach assumes conditional independence across classes, treating each class separately. Instead, one can employ contrastive learning for feature distillation, transferring all the information in the teacher's representation. Examples include model compression and cross-modal knowledge transfer. Finally, contrastive learning can be extended to more than two views, or from views coming from paired modalities [259].

1.4.2 Paired multi-modal contrastive learning

Images paired with open-ended text descriptions have emerged as a pretraining strategy. Natural language is a weaker form of supervision, and thus it is not limited by a finite set of predefined categories. In contrastive language-image pretraining (CLIP) [214], we consider a mini-batch of size B consisting of aligned image-text pairs $[(v_1, t_1), \ldots, (v_B, t_B)]$. Each modality is encoded using a separate encoder:

$$z_{\text{img}} = g_{\text{img}}(v_i), \quad z_{\text{text}} = g_{\text{text}}(t_i), \quad z_{\text{img}}, z_{\text{text}} \in \mathbb{R}^d. \tag{1.31}$$

Text is typically represented by representing each word as a d-dimensional vector that is learned during training. The critic is defined as in Eq. (1.17), yielding ℓ_2-normalized vectors. By treating one modality as anchor and the other as target, the contrastive objective encourages samples from one modality to match only their corresponding positive pairs, while pushing them apart from the remaining $k = B - 1$ negative samples in the mini-batch. For instance, using the image modality as an anchor, we can compute the loss for a positive pair (v_i, t_i)

$$\mathcal{L}_{\text{contrast}}^{\text{img}\to\text{text}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{h_\theta(v_i, t_i)}{\sum_{k=1}^{B} h_\theta(v_i, t_k)} \tag{1.32}$$

Similarly to Eq. (1.32), one can define $\mathcal{L}_{\text{contrast}}^{\text{text}\to\text{img}}$, and the final loss for contrastive vision-language pretraining at each iteration can be computed as $\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{\text{contrast}}^{\text{img}\to\text{text}} + \mathcal{L}_{\text{contrast}}^{\text{text}\to\text{img}})$. After training, given a set of text candidates, top-1 image-text similarity is used to predict the image label. While web-scale datasets used in pre-training partially overlap with current generalization benchmarks, it is insufficient to explain CLIP's superior generalization performance on current benchmarks [186]. Finally, similar representations to CLIP can be obtained with image captioning [266]. Several works have indicated that high-quality dataset creation (e.g. descriptive captions, high image diversity) is the key to acquiring robust vision-language representations [81, 82, 29, 293].

Multi-modal contrastive tuning. Apart from representation learning (starting from random initialization), Eq. (1.32) can teach a text encoder to match representations from an already trained visual encoder [306]. The weights of the image encoder are not updated during contrastive tuning. This results in new vision-language capabilities for downstream tasks such as zero-shot classification or retrieval. Similarly, natural images can serve as the core representation to facilitate switching between multiple modalities [91]. Starting from a pretrained image and text encoder (i.e. CLIP), it is possible to learn a joint representation across six different modalities (images, text, audio, depth,

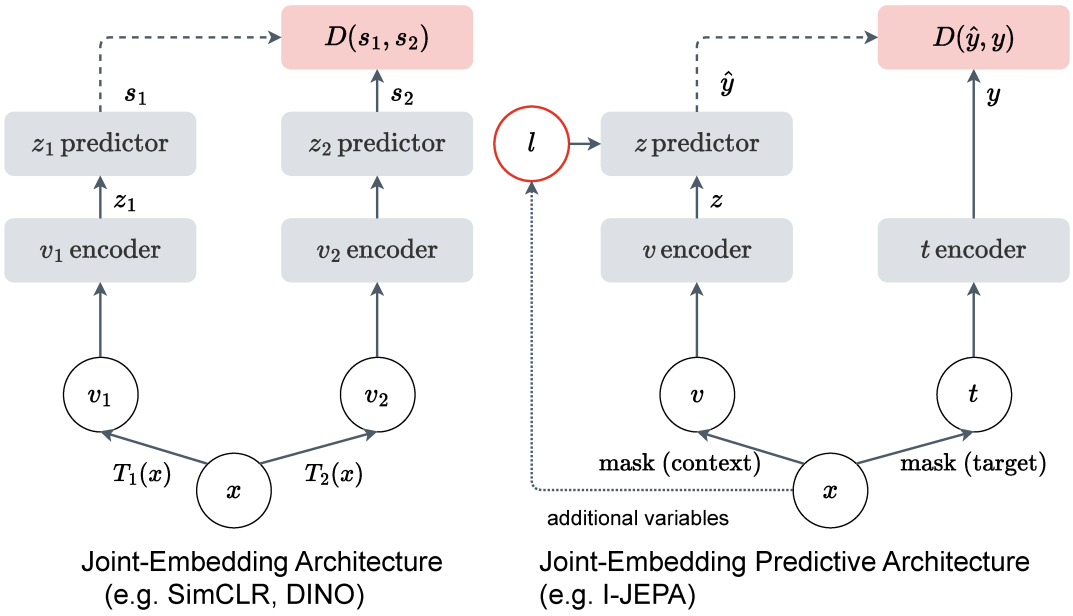


Figure 1.4: **Joint-embedding architectures for self-supervised learning, in which the system learns to capture the relationships between its inputs.**

thermal, and IMU data). Recently, Jose et al. [133] showed that additional vision layers after the pre-trained vision encoder are beneficial for contrastive tuning.

1.5 Joint-Embedding Self-Supervised Approaches

In contrastive learning, a large batch size is required [288] to ensure a sufficient number of negative samples. Using a large batch size poses computational challenges at scale. Alternative self-supervised methods attempt to avoid this bottleneck by using negative samples in more implicit and indirect ways [317]. The purpose of negative samples is to prevent representation collapse [42] when dealing with smaller batch sizes. To do so, such methods often require training additional components and hyperparameters that are carefully tuned to avoid collapse, making them difficult to train. We will broadly refer to this category of methods as *joint-embedding architectures (JEA)*.

In Fig. 1.4, we first present a high-level overview of common self-supervised learning strategies. SSL methods, including JEA, exhibit desirable properties, such as robustness to distribution shifts [202], and provide strong global and local features [240]. In this work, we group DINO, MSN, iBoT, and DINOv2 as JEA methods [42, 317, 202, 14, 16]. A common improvement to these approaches is processing multiple views per sample at each iteration, which can become costly at large scales. Recently, a newer category of methods aims to combine the best of JEA and reconstruction methods called joint-embeddings *predictive* architectures (JEPAs) [16]. In this direction, they only use input masking and make predictions in a representation space. The input masking designs global crops of an image with small missing patches (the context), which are predicted in representation space, based on a second model that processes these patches. We give a specific example in Section 1.6.

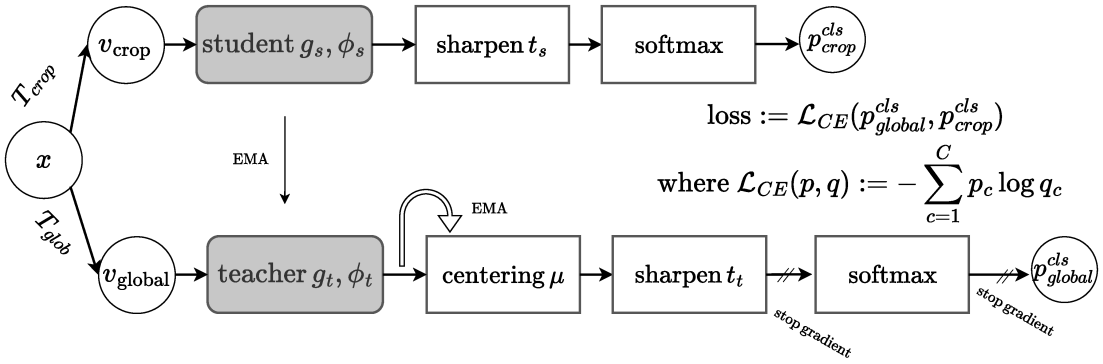


Figure 1.5: **Overview of the DINO framework: self-distillation without labels.** Two different random transformations of an input image x are passed to the student and teacher networks. The output of the teacher network is centered with a mean μ calculated as an exponential moving average (EMA) over the mini-batch. Each network outputs an C-dimensional vector, normalized with a temperature-scaled softmax. Gradients are propagated only through the student branch. The teacher parameters are updated with an EMA of the student parameters.

1.5.1 DINO and self-distillation

The DINO framework [42] uses two encoders: i) a so-called teacher network $g_t(.)$ parameterized by θ_t and a student network $g_s(.)$ parameterized by θ_s. During pretraining, the cross-entropy loss encourages the student's representation to align with the teacher's representation. In contrast, the teacher is updated using the student's weights as shown in Eq. (1.18) [51]. This paradigm, known as self-distillation, assumes that both models have the same architecture.

During training, each image $x \in \mathbb{R}^{3\times H\times W}$ is processed as follows. We construct different local and global crops v_{crop} for each image with a multi-crop strategy. We initially present the DINO framework for a single crop $v_{\text{crop}} = \mathcal{T}_{\text{crop}}(x)$ and a global view $v_{\text{global}} = \mathcal{T}_{\text{global}}(x)$. Importantly, the view v_{crop} can either be a "global view" (i.e. a large crop) or a "local view" (i.e., a small crop). The first view v_{crop} is fed to the student network, while the global view v_{global} is fed to the teacher network:

$$z^{cls}_{\text{crop}} = g_s(v_{\text{crop}}|\theta_s, \mathcal{T}_{crop}) = g_s(v_{\text{crop}}) \text{ and similarly } \; z^{cls}_{\text{global}} = g_t(v_{\text{global}}), \tag{1.33}$$

where the *cls* refers to an aggregate representation of the network, such as the added classification token in a Vision Transformer (ViT) architecture [70]. For the sake of completeness, when a view of arbitrary spatial dimensions is fed through the network, it receives an interpolated positional encoding based on its length. The number of patches L is determined based on the patch size and the spatial dimensions H, W as $L = \left\lfloor \frac{H\times W}{\text{patch}^2} \right\rfloor$. The total sequence length L_{all} adds all special tokens to L, such as the *cls* token.

In the *multi-crop* setup, all crops v_{crop} are passed through the student network. The global views v_{global} are passed through the teacher, encouraging "local-to-global" correspondences. Moreover, the DINO framework adds the following processing steps to prevent collapse:

- Adds weight-normalized linear layers [231] in the last layers called the "DINO

heads", $\phi_s(.), \phi_t(.) : \mathbb{R}^d \to \mathbb{R}^C$, placed after the student and teacher networks respectively.

- Centers the teacher-computed features using a vector $\mu \in \mathbb{R}^C$, which implicitly uses negative samples from the mini-batch and pushes their features apart.
- Computes a temperature-weighted softmax score of both features in $\mathbb{R}^C$, which they then can compare using cross-entropy. The resulting probability vectors are sometimes called *prototypes*, clusters, or soft-classes.

Mathematically, the steps to get probability vectors for each view are as follows:

$$p^{cls}_{\text{crop}} = \text{softmax}(\phi_s(z^{cls}_{\text{crop}})/\tau_s), \in \mathbb{R}^C \tag{1.34}$$

$$p^{cls}_{\text{global}} = \text{softmax}((\phi_t(z^{cls}_{\text{global}}) - \mu)/\tau_t), \in \mathbb{R}^C \tag{1.35}$$

where $\tau_s, \tau_t > 0$ is the temperature parameters. In practice, $\tau_t \leq \tau_s$, making the predictions of the teacher more confident than the student. Intuitively, centering aims to avoid the trivial solution of mapping all vectors to a single class, while temperature weighting is introduced to avoid a uniform distribution of p^{cls}_{global} among C. The loss function d_{CE} encourages p^{cls}_{crop} to match the teacher predictions p^{cls}_{global}.

For two global views and l local views, the per-sample loss computes the average cross-entropy over all $2 + 2l$ valid pairs of views per image:

$$L_{\text{DINO}} = \frac{1}{2+2l} \sum_{i=1}^{2} \sum_{j=1}^{l+2} \mathbb{1}_{[j \neq i]} \mathcal{L}_{\text{ce}}(p^{(i)}_{\text{global}}, p^{(j)}_{\text{crop}}), \tag{1.36}$$

where the indices i, j run over different views. L_{DINO} is then averaged for all images B in the mini-batch. During training, the student weights θ_s and head parameters of $\phi_s(.)$ are updated using stochastic gradient descent-type iterative optimization. During the gradient computation, the teacher parameters and head are constant (not optimized). After the student's parameter update, the teacher parameters θ_t and the ones corresponding to the head ϕ_t are updated via exponentially moving averages (EMAs) of the student. Additionally, the centering vector μ is updated via an EMA of the teacher's predictions. The three EMA-related hyperparameters and the temperatures can change along the optimization trajectory, and their schedules are design decisions that impact convergence.

Importantly, the balance between the centering operation Eq. (1.35), and the choice of temperatures τ_s, τ_t is hard to achieve as highlighted by Caron et al. [42]. A large number of concurrent studies aim to address this issue by increasing training stability without requiring as many post-hoc hyperparameter tweaks [41, 289, 94, 125]. After training, the student network is discarded. Finally, the teacher network $g_t(.)$ with parameters θ_t is used for evaluation.

Improvements on the DINO framework. The current focus is on developing training strategies that are simpler [289], accelerate convergence, and ideally enhance robustness to distribution shifts. For instance, Zhang et al. [310] downsample images in the early training stage to encourage the model to learn low-frequency features, which accelerates convergence. Wu et al. [289] have significantly simplified the pipeline by removing the predictors $\phi(.)$ and directly minimizing the Euclidean distance between $z^{cls}_{\text{crop}}, z^{cls}_{\text{global}}$. To avoid collapse, the authors [289] directly penalize the covariance of

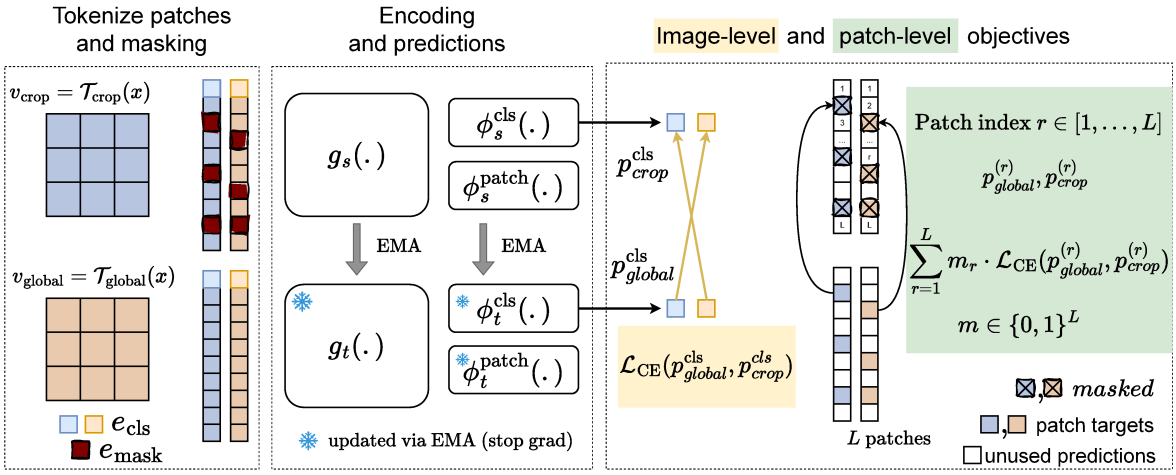


Figure 1.6: **Overview of DINOv2 framework**. DINOv2 combines masked image modeling (patch-level objective) with the self-distillation objective of DINO (image-level). Given two global views $v_{\text{crop}}, v_{\text{global}}$ of an image x, both views are passed through a teacher and student encoder. The cls token outputs across *different* views are used to compute cross-entropy. The green area illustrates the patch-level prediction objective between *tokens from the same view*. The superscript r is used for indexing the L patch outputs. A percentage α of masked tokens are masked and replaced by e_{mask} for the student. The following are **not** shown for clarity: the KoLeo regularization, the Sinkhorn-Knopp centering applied after $\phi_t^{cls}(.)$.

the features as a regularization strategy [304]. Other variations replace the centering operation with the Sinkhorn-Knopp (SK) algorithm, which is typically run for a few iterations [41, 14, 94]. To encourage the model to utilize the full set of prototypes, the mean entropy maximization regularizer is introduced, similar to [268, 14, 15].

1.5.2 DINOv2

Apart from the image-level objective of DINO, the pipeline of DINOv2 [202], based on iBOT [317], additionally learns an auxiliary patch-based objective. However, this only applies to ViTs. The learned patch-based representations contain fine-grained semantic information about the local neighborhood of the patch (e.g. 16×16 for an image of size 224×224) that is beneficial for dense vision tasks.

To incorporate the patch-level objective, we randomly mask a percentage $\alpha \in [0, 1]$ of the L input patches that are passed to the student encoder, but not to the teacher. The masked tokens are replaced with a learnable masked token denoted as e_{mask}. The sequence mask is denoted as $m \in \{0, 1\}^L$, where one corresponds to a masked patch. Here, we use the index r when indexing patches to differentiate from the notation used in DINO that indexed views. The idea of a shared masked token across all views and images in the mini-batch is conceptually similar to the masked autoencoder [100]. Notable additional changes from DINO include:

1. **Extra prediction heads**: An additional predictor $\phi^{patch}(.)$, also known as a head, is processing the patch output representations of the encoder (for both teacher and student networks). $\phi^{patch}(.)$ is shared for all patch representations.

2. **Centering**: Instead of the centering operation, the Sinkhorn-Knopp (SK) algorithm [57, 41, 14, 94] is applied for both the *cls* output and the patch-based representations of the teacher. SK uses features from all images in each mini-batch.

3. **KoLeo regularization.** In DINOv2, the *KoLeo* regularizer is applied to a global view z^{cls}_{crop} of the student network. The regularizer is derived from the Kozachenko-Leonenko differential entropy estimation [61]. It encourages a uniform span of the features within a mini-batch:

$$L_{\text{koleo}} = -\frac{1}{B}\sum_{i=1}^{B} \log \max_{j \neq i} \text{sim}(z_i, z_j), \quad \text{where } j \in \{1, \ldots, B\}, \tag{1.37}$$

and $\text{sim}(z_i, z_j) = \frac{z_i^T z_j}{||z_i||\ ||z_j||}$. Excluding self-similarities, nearest neighbors within the same mini-batch are pushed further apart. Nonetheless, Eq. (1.37) requires a sufficiently large mini-batch size to converge [289].

The patch-level objective for the two global views that are processed from the teacher and student network is $L_{\text{iBOT}} = \sum_{r=1}^{L} m_r \cdot \mathcal{L}_{\text{CE}}(p^{(r)}_{global}, p^{(r)}_{crop})$, where r indexes patches. The final loss at each iteration combines all three losses:

$$L_{\text{DINOv2}} = L_{\text{DINO}} + L_{\text{iBOT}} + \lambda L_{\text{koleo}}, \quad \lambda > 0, \tag{1.38}$$

where a common value for λ is 0.1.

Alternative regularization strategies of JEA. Follow-up works proposed alternative regularization terms. Instead of spreading the features z^{cls}_{crop}, Govindarajan et al. [94] propose the KoLeo regularization to distribute the prototypes p^{cls}_{crop} uniformly. Another established regularization objective is the mean entropy maximization (ME-MAX) regularizer, also used in [15, 268, 14], defined as:

$$L_{\text{ME-MAX}} = -H(\bar{p}) = \sum_{c=1}^{C} \bar{p}_c \log(\bar{p}_c), \text{ where } \bar{p} = \frac{1}{Bl}\sum_{i=1}^{B}\sum_{j=1}^{l}(p^{cls}_{\text{crop}})_{ij}, \tag{1.39}$$

where l is the number of views that are passed in the student network, and B is the mini-batch size. Wu et al. [289] directly penalize the covariance of the features to avoid representation collapse. For a hyperparameter $\delta > 0$, they utilize an approximation of the rate distortion of a Gaussian random variable with covariance $\mathbf{\Gamma} = \frac{1}{B-1}\sum_{i=1}^{B}(z_i - \bar{z})(z_i - \bar{z})^T \in \mathbb{R}^{d\times d}$ where $\bar{z} = \frac{1}{B}\sum_{i=1}^{B} z_i$ and an identity matrix $\mathbf{I}_d \in \mathbb{R}^{d\times d}$, defined as:

$$L_{\text{rate}} = -\frac{1}{2}\log\det\left(\mathbf{I}_d + \frac{d}{\delta^2}\mathbf{\Gamma}\right) \tag{1.40}$$

In practice, the covariance matrix $\mathbf{\Gamma}$ is calculated per global view from the student features and then averaged over all global views. Notably, Eq. (1.40) yields a measure of the size of the covariance, even if the underlying variables are non-Gaussian [289]. A conceptually similar formulation has been applied in [305, 21]. Finally, when dealing with pseudo-labels/prototypes in the pretraining objective (i.e. cluster indices, sample index in the dataset), label smoothing has been utilized as a regularizer [125].

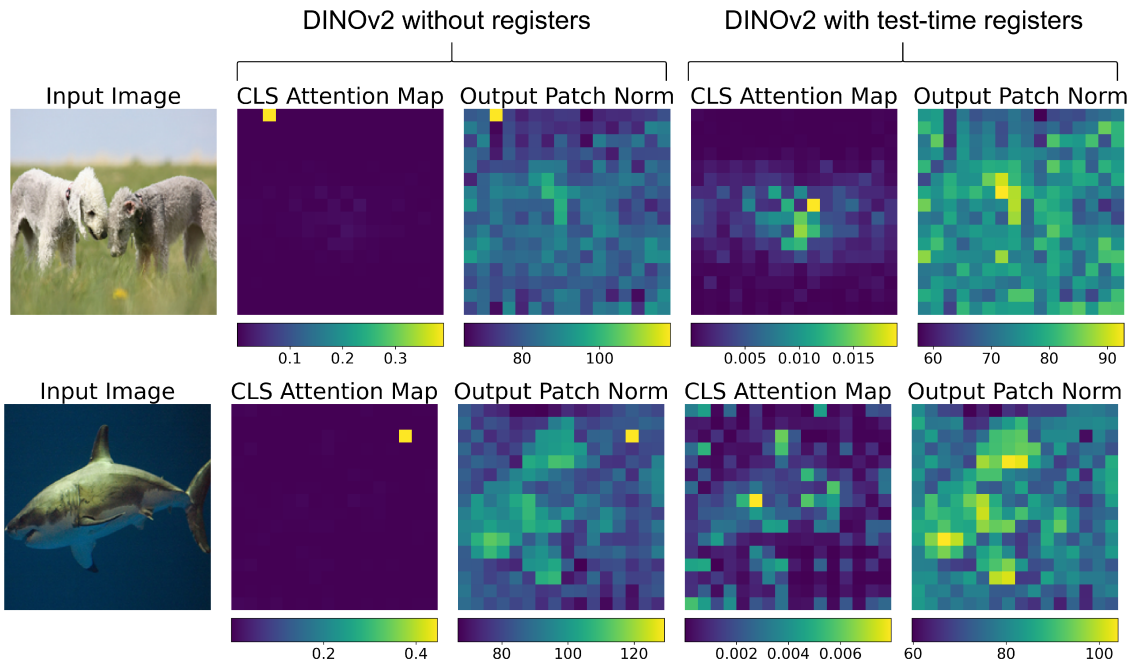


Figure 1.7: **DINOv2 attention artifacts can be addressed with registers**. Registers are additional tokens that are appended to the input sequence of the Vision Transformer. [59]. Each row represents inference using the image shown in the first column. The mean CLS attention map is shown, upsampled to the image dimensions (*second and fourth column*). Artifacts are high-norm tokens (*third column*) appearing in uninformative background areas (*second column*). In the last two columns, we perform inference using the latest training-free approach of test-time registers using the official implementation from Jiang et al. [130].

1.5.3 Notable developments based on DINOv2

One issue with the additional objectives of DINOv2 is the noisy attention maps. The attention maps are usually upsampled to the image dimensions using interpolation and used as label-free segmentation masks. The artifacts typically correspond to a sparse set of high-norm tokens appearing during inference in low-informative background areas of images (outlier tokens), as shown in Fig. 1.7, second and third columns. This behavior can also occur in vision-language models, such as CLIP [130]. To reduce artifacts in the attention maps, Darcet et al. [59] propose a simple solution based on providing additional tokens to the input sequence during training, called *registers*. To remove artifacts without retraining, Jiang et al. [130] propose shifting the high-norm activations into an additional untrained token, mimicking the behavior of registers (Fig. 1.7).

Another regularization technique introduced in DINOv3 [240] is Gram anchoring, which addresses the collapse of dense features that occurs with large models trained over long schedules. This regularization objective aims to preserve the quality of dense feature maps by enforcing the Gram matrix (the matrix of all pairwise dot products of patch features) of the student network to align with that of an early iteration of the teacher network) [240]. Other significant hurdles exist in applying visual-only pre-training in different domains with varying input channel dimensions. Such domains

with inconsistent input channel properties include microscopy and satellite images. To address this, Lorenci et al. [179] train DINOv2 on individual randomly drawn channels, ignoring any cross-channel dependency in the learning process. Other recent post-training practices involve:

1. **Post-training with higher resolution images** [264, 237, 31]. To encounter varying test-time image resolutions, Simeoni et al. [240] utilize mixed resolutions of global and local crops.

2. **Distillation**, where the largest-scale teacher model is distilled into a series of smaller models to fit computational demands. This can be achieved with minor modifications of the same framework. During the distillation process, the weights of the teacher are unchanged, and only the smaller-sized student network(s) are being updated with gradient descent.

3. **Language compatibility**. Vision-only models can acquire text-to-image capabilities by training a text encoder to align with the "CLS" (classification) token output of a frozen vision model [306]. Jose et al. [134] recently showed that concatenating the CLS token representation with the average patch representation yields state-of-the-art results both in zero-shot classification and semantic segmentation. Using the open-source weights (DINOv3 ViT-L + text encoder), the author was able to reproduce the zero-shot accuracy on ImageNet1K of 82.68% and 81.42% with and without prompt engineering, respectively.

1.5.4 Large-scale data collection and curation for self-supervised learning

In both vision and language models, training on arbitrary collections of web images or enormous pools of ill-curated text corpora has led to subpar performance on standard benchmarks [258, 286, 55]. This poor performance is mainly attributed to the imbalance of uncurated data [176]. Visual concepts or the distribution of languages tends to be highly non-uniform, leading to biases toward the overrepresented concepts or data in the learned representations [283, 273]. In the unsupervised setup, we abstractly refer to the data belonging to "concepts". A visual *concept* is an abstract term that emerges from the data. Vo et al. [273] describe the concept as the shared content of a group of samples that are similar according to human perception.

Selecting samples from the unlabelled, uncurated data for scaling up is coined as *automatic data curation*. An ideal dataset is sufficiently large, and provides a diverse and balanced coverage of visual concepts [135, 115, 219] and leads to more robust features [95]. The naive idea of collecting and combining existing curated datasets does not necessarily translate into better downstream performance. Given a fixed training budget and a self-supervised model trained on a "seed" visual dataset, such as ImageNet-21K [226], one can use this model to curate samples. We refer to the initial training data as the "seed data". Similarly, LLMs are typically anchored around high-quality data sources such as Wikipedia [263]. Below, we describe different approaches to enlarge the seed data. Importantly, the uncurated data first requires careful deduplication [209]. Dominant strategies for automatic data filtering can be categorized into:

1. **Retrieval-based dataset augmentation [213, 295, 202]**: For each image in the seed data, one can select the κ most similar samples from the uncurated collection [202]. This effectively multiplies the number of training data by κ. Cosine similarity between images is commonly used.

2. **Clustering-based**: To increase diversity, feature-based clustering on the uncurated data can be applied, such as k-means [178]. This is particularly beneficial for small-scale datasets. After clustering, for each seed image, κ images from the corresponding cluster are sampled. However, selecting images based on seed samples from predefined datasets may ignore a plethora of visual concepts in web-based collections [273]. In [273], the authors aim to create a balanced subset of visual concepts by clustering the uncurated data in a near uniform way. To achieve this, they develop a hierarchical k-means variant where only the centroids are clustered after initially running vanilla k-means.

3. **A mixture of retrieval and clustering** [202], given a set of seed datasets, can also be applied. In [202], Oquab et al. argue that clustering is beneficial for small seed datasets and retrieval is used for the large ones, like ImageNet-1K.

1.6 Image Representations *Without* Hand-crafted Augmentations

Theoretical studies such as Zhang and Ma [312] mention that: "*The most indispensable process in contrastive learning is the data augmentation module*". To understand contrastive learning, Saunshi et al. [232] view data augmentations as inductive biases and develop theorems valid in the case of linear representations. However, the majority of visual JEA were developed on ImageNet, which might affect the derived conclusions on the necessity of handcrafted data augmentations.

Another intuitive way to assess what type of invariances are learned is to train a generative model conditioned on the learned feature representations [35] instead of class indices. In Fig. 1.8, we visualize generated samples based on trained diffusion models from Bordes et al. [35]. The representation at the encoder level does not allow as much variance in the generated samples as when using the outputs of the head. For instance, information about the pose and size of the animal is stored in the encoder's representation. The observed variance when the head features are used indicates significant information loss.

JEA. The insofar adopted augmentations (Fig. 1.3) include resized crops of varying scales, horizontal flips, color jittering (such as brightness changes), greyscale conversions, and Gaussian blurring. Each transformation is randomly applied with some probability. Nonetheless, when the training data is large enough, strong representations can be obtained only using random cropping [196]. Simpler pipelines are easier to generalize to other types of data, such as graphs, 3D medical images, and time series. To this end, starting from DINOv2 as baseline, Moutakanni et al. [196] only apply cropping without random resizing and random patch masking with comparable results at large-scale datasets. For instance, given a 256×256 image, global crops are taken with a fixed aspect ratio of 224×224, and local crops are taken at 98×98 resolution [196].

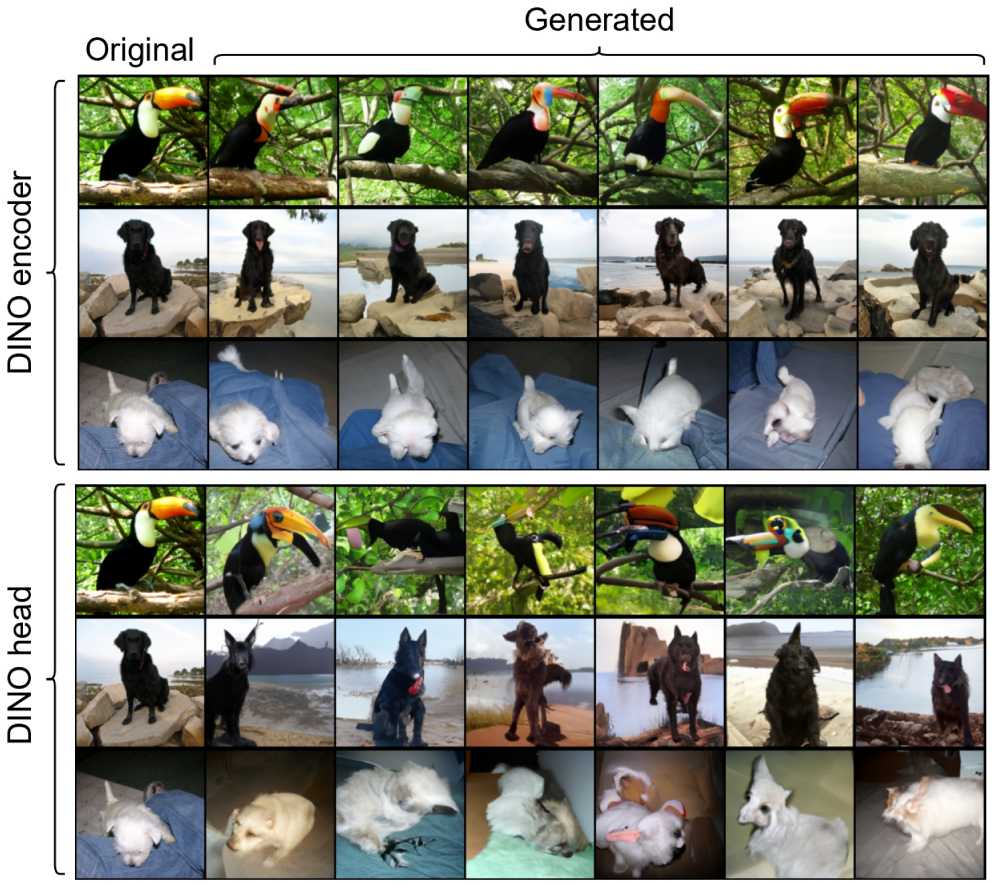


Figure 1.8: **Generative learning conditioned on feature representations based on the image shown in the first column.** The developed framework, called RCDM [35], is a diffusion-based model conditioned on DINO ResNet50 feature of the encoder (first three rows) or the head/predictor (last three rows). Samples are produced by the author based on the official implementation [35] using ImageNet test samples for conditioning.

JEPA. A newer alternative approach is joint-embeddings *predictive* architectures (JEPAs) [16]. JEPAs are non-generative and do not rely on custom augmentations. The predictions are made in a representation space similar to DINOv2. A famous exemplar method [16] is shown in Fig. 1.9: four target blocks in pixel space are randomly sampled with a ratio between 0.15 and 0.2 of the original image. Similar to a global view, we randomly sample a context block with scale in the range (0.85, 1.0) and remove any overlapping target blocks [16]. The context is processed by the student network, which is updated using stochastic gradient descent, while the teacher processes the target blocks. The loss is simply the average L2 distance between the predicted patch-level representations (student's predictions) and the target patch-level representation (teacher's predictions).

1.7 Representation Convergence

While this manuscript primarily covers image representations, there has been an increased interest in discovering shared patterns among different representations [123]. Similarity

Figure 1.9: **An illustration of the joint-embeddings *predictive* architecture's (JEPA) sampling strategy.** Given an original image, we sample one context block (second column) mask with random scale in the range (0.85, 1.0), shown in **black color**) and four possibly overlapping target block masks (shown in **grey color**) with random scale i.e. 15%. Any region in the context block mask that overlaps with any of the four target blocks is removed from the context image. The sample is taken from the STL10 dataset.

between representations can be investigated between i) training on a single modality using a different objective (i.e. supervised versus self-supervised) [71], ii) a unimodal representation with a multi-modal representation (vision versus vision-language models), or even c) different modalities with different pipelines (a vision model with a language model) [123]. Existing works [19, 89] uncover that representations from well-trained self-supervised vision models align with their supervised counterparts. Instead of comparing features, Dravid et al. [71] studied individual neurons (activations), and found "Rosetta Neurons" that are activated by the same pattern, discovered independently, across several vision models.

While the results depend on the measure of alignment, networks that generalize better tend to converge to more similar representations than networks that memorize, even on small-scale data such as CIFAR10 [195]. An example of memorization in this context is fitting labels that have been randomly permutated [308]. Huh et al. [123] find that among 78 vision models, the ones with high transfer performance form a tightly clustered set of representations, while models with weak performance have more variable representations. This finding suggests that well-performing models represent information similarly, even if they have different architectures and training objectives.

Representations are converging across modalities. Recently, in multilingual settings for vision-language learning, there have been cross-lingual performance synergies on billion-scale datasets [55]. When metadata, data curation, model capacity, and training are scaled jointly, English and non-English data become mutually beneficial, breaking the curse of multilingualism. The infamous GPT4 technical report [1] demonstrates that jointly training a language model with a vision model enhances performance on language tasks compared to training the language model in isolation. Sharma et al. [236] assess the visual understanding of LLMs (trained only on language data) by converting images into code. Subsequently, the authors train a vision system using synthetic LLM-generated code samples (text) converted back into images. Thus, LLMs possess a rich understanding of visual structures, which supports potential performance synergies.

To measure cross-modal alignment, Huh et al. [123] utilize paired datasets, such as the Wikipedia captions dataset [250]. By comparing the similarity between two images (an image depicting an apple and an orange) and their corresponding texts, the authors

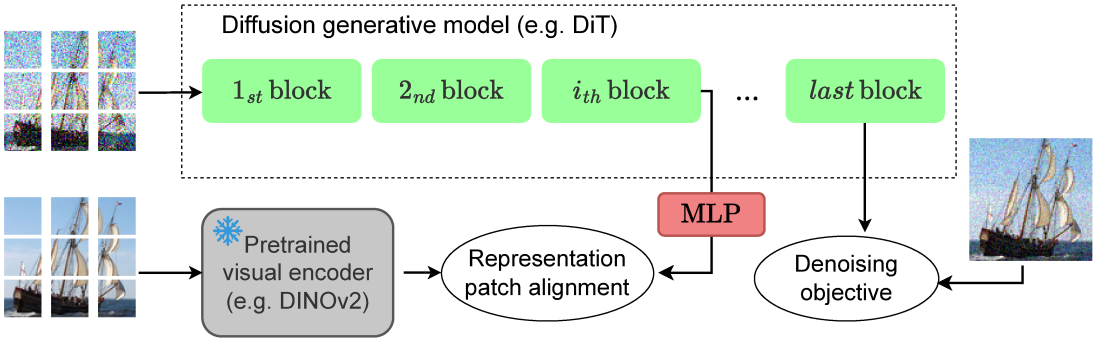


Figure 1.10: **Overview of the REPA [302] framework.** During training of the diffusion model, such as DiT [206], clean patch representations from a trained visual encoder are aligned with noisy intermediate representations of the diffusion model. A small multi-layer perception (MLP) is used to match the per-patch dimensions. Cosine similarity is maximized between patch representations.

find that the better an LLM is at language modeling, the more it tends to align with vision models, and vice versa [123]. As alignment improves with superior language/vision models, superior downstream performance has been observed so far. Possible explanations include, but are not limited to, the simplicity bias described by [123]: "Deep networks are biased toward *finding simple fits to the data, and the bigger the model, the stronger the bias*. Therefore, as models get bigger, we should expect convergence to a smaller solution space."

1.8 Generative Learning using Visual Representations

It is argued that training of generative models is bottlenecked by the lack of representation learning [302]. In REPA [302], the authors guide the latent diffusion generative model with external visual representations (i.e. DINO features), as illustrated Fig. 1.10, with significant improvements in image fidelity. Similar efforts [303] match low-level layers with a single high-level class (CLS) token of the clean representation for denoising. Recent work aligns pretrained visual encoders with the generative variational autoencoder used in latent generative modeling [294]. Kouzelis et al. [150] fuse the principal components of the patch representations from DINOv2 with the VAE latents (typically four channels) as inputs to the denoising diffusion model. Fusion is realized as a channel-wise summation of the two token representations after spatial interpolation and channel matching. Channel matching is realized as a combination of PCA and a trainable projection matrix for efficiency reasons [150]. Wang & He [13] demonstrate a similar result by spreading the intermediate noisy features without a pretrained visual encoder.

An alternative to aligning visual representations to regularize generative models is conditioning. This category of methods aims to improve unconditional generation, which currently lags significantly behind conditional generation in the literature. Features from real training images cannot be used directly for diverse generative learning apart from visualization purposes, as described in RCDM [35] (Fig. 1.8). Adaloglou et al. [2]

use the DINO features for clustering and condition the diffusion model on the cluster assignments. Nonetheless, parametric clustering requires selecting the number of clusters, which is challenging for arbitrary datasets. Another approach involves conditioning with the nearest neighbors in the feature space of real images, parallel to image features [43]. Li et al. [165] decompose the generative learning using visual representations in a two-stage framework. First, the authors train a representation-generative model that samples from a noise distribution and generates a representation that aligns with the visual encoder. Then, it learns an image generator that maps a noise distribution to the image distribution, conditioned on the sampled representation from the previous model.

To summarize, the plethora of recent approaches [2, 165, 161, 238, 32] highlight the usefulness of representation learning as a first step for visual generative learning.

2. Downstream Tasks that Leverage Visual Representations

2.1 Visual Out-of-Distribution Detection

2.1.1 Introduction and taxonomy of tasks

A trustworthy visual recognition system should not only provide accurate predictions, but also reliably detect anomalous samples [297]. This capability is essential in real-world applications where safety and reliability are non-negotiable. Blind reliance on the predictions of neural networks can lead to undesirable or even dangerous behavior. For instance, an autonomous driving system operating on a foggy day should alert the driver and hand over control. Once a sample is identified as OOD, the associated predictions must be treated as unreliable [296, 311]. To mitigate such risks, models are required to provide an estimate of certainty in their predictions [23], commonly referred to as confidence [106].

Confidence in neural networks is defined as the quantifiable measure of certainty that a given sample's prediction is correct. A model is considered calibrated when its confidence estimates align with its true predictive accuracy. For example, if an algorithm classifies a set of images with a confidence interval of $[95\%, 100\%)$, we expect at least 95% of those class predictions to match the ground truth labels. Detecting anomalies is therefore crucial to prevent models from making potentially uncalibrated predictions on unseen data [11, 222, 97]. Despite recent progress, neural networks are known to exhibit overconfidence [199], or sometimes assigning higher likelihoods to anomalous samples than to the training data [198], which follows the in-distribution (ID). Detecting samples that lead to miscalibration naturally falls into the anomaly detection task.

Anomaly detection (AD) [46] is the most widely used framework for identifying deviations from the in-distribution and encompasses most other categories. AD methods are inductive, meaning that, given a set of training samples, we aim to make decisions about unseen ones. Formally, AD aims to detect any anomalous samples that deviate from the predefined normality during testing [46]. The deviation can happen due to a particular type of in-distribution shift. Importantly, anomaly detection does not assume semantic categories for the in-distribution (ID), nor does it require semantic differentiation in the ID samples. For this reason, we present more precisely defined tasks below.

1. **Outlier detection** [25, 169] aims to detect anomalous samples from an observation set, which are markedly different from the others, due to a distribution shift. Unlike AD, this task is transductive, meaning that the observation set contains anomalous samples, referred to as outliers.

2. **Novelty detection** [255, 177] aims to detect any test samples that do not fall into any training semantic category. It is essential to note that the primary objective of novelty detection is to distinguish novel samples from known ones, without requiring classification among the ID classes. A special case includes one-class novelty detection, where samples of a single class are considered as the ID, and the goal is to reject all other classes in the dataset. We abstain from using this term, as it overlaps with out-of-distribution detection. Novelty detection and outlier detection have become less active in the era of large-scale vision-language models [192].

3. **Sensory (anomaly) detection** [128] aims at detecting non-semantic shifts. Non-semantic shifts, also known as covariate shifts, occur when the image undergoes a shift but the underlying label remains unchanged. In natural image classification, examples of covariate shifts can arise from the camera's angle, adversarial examples [93], style changes [104, 90], weather conditions [105], or generally anomalies within images [39]. Other use cases include classifying defects in industrial manufacturing [128], which typically have a wide range of variations. Sensory detection remains relevant and challenging, even with the advent of vision-language models, as evidenced by industrial anomaly detection benchmarks such as MVTec-AD [28] and VisA [322]. Finally, sensory anomaly detection can be extended to anomaly segmentation, where the aim is to localize defects on a pixel level.

4. **Semantic out-of-distribution (OOD) detection** detects test samples drawn from a distribution that is different from the training distribution. The definition of distribution needs to be well-defined according to the application in the target, and often refers to the semantic categories. In this case, OOD samples should not have overlapping labels w.r.t the training ID data. OOD detection, also known as open set recognition [233], is similar to multi-class novelty detection. The core difference between novelty detection is that OOD detection further assumes the presence of an ID classifier, whose performance (i.e. ID accuracy) needs to be maintained by rejecting semantically shifted samples. Finally, OOD detection can be defined more broadly by revisiting the definition of the normal/alternative distribution [297]. In this manuscript, we focus on distinguishing semantic categories.

5. **Full-spectrum OOD detection (FD-OOD)** [298]. Whilst an OOD detector needs to be sensitive to semantic shifts, covariate shifts should not be flagged as OOD (Fig. 2.1). The above balance of sensitivities to distribution shifts is referred to as OOD robustness or FS-OOD detection, and is a crucial step towards generalization [257, 220]. Existing works study these two tasks (sensory detection, semantic detection) independently. ID generalization further measures the performance change (e.g. accuracy drop) on the covariate samples versus the ID test set. As carefully-curated OOD benchmarks are developed with minimal semantic overlap [33], FS-OOD receives increased attention [311]. Recent work has shown that even with access to labeled ID data, OOD algorithms remain vulnerable to covariate shifts [299]. FS-OOD detection remains a critical challenge for future OOD methods.

Other less relevant topics, which are outside the scope of this work and will not be discussed, include learning with rejection, domain adaptation [278], novelty discovery [98],

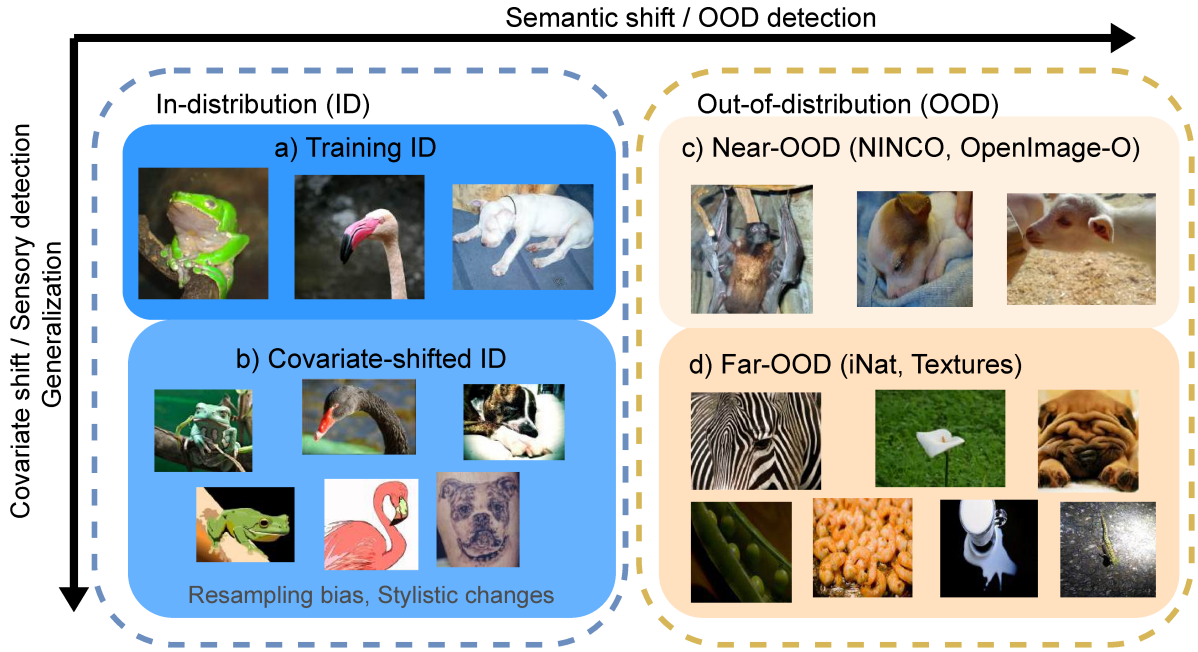


Figure 2.1: **An visual illustration of full-spectrum OOD detection.** We use ImageNet as the in-distribution (ID). An ideal OOD detector should be sensitive to and reject semantically shifted samples and, at the same time, robust to covariate shifts [296].

open-world recognition [176], and conformal prediction. The reader is encouraged to read Yang et al. [297] for more details. In this work, we focus on semantic out-of-distribution (OOD) detection and FS-OOD detection.

2.1.2 Out-of-Distribution Detection Methods

OOD methodologies. Following Yang et al. [297], we categorize OOD methodologies into four groups: 1) classification-based methods, which primarily rely on an ID classifier, 2) density-based methods [207, 37], which model the data distribution and flag samples in low-density regions as OOD, 3) distance-based methods, which identify OOD samples in feature space by measuring their distance from class centroids or prototypes, and 4) reconstruction-based methods [163, 73, 154, 175]. The intuition behind reconstruction approaches is that a generative model trained on the ID data usually fails to reconstruct OOD samples as effectively as ID samples, thereby indicating anomalies [64]. A more fine-grained task-oriented categorization can be found in [180]. In this work, we focus on *post-hoc methods*, which can be classification or feature-based distance methods. Critically, post-hoc or post-training approaches do *not* require training modifications, aligning with the use of general-purpose visual representations for downstream OOD detection.

OOD detection methods with labeled ID samples.

Consider a real-world deployment scenario, such as image classification, where the classifier has been trained or fine-tuned on the ID. Alternatively, when an ID classifier is not available, zero-shot inference using CLIP can be employed by utilizing the ID label names. The task of out-of-distribution (OOD) detection is to determine whether a new test

sample is drawn from the ID or not. Rejected OOD samples can be handed over to human reviewers. The decision-making process in OOD detection can be formalized by an OOD score function, $s(x)$, which maps an input sample x to a scalar value. We define this score such that higher values indicate higher confidence that the sample is ID. The sample x is classified as ID if its score $s(x)$ is greater than or equal to a predetermined threshold ξ:

$$x \in \begin{cases} \text{ID}, & \text{if } s(x) \geq \xi \\ \text{OOD}, & \text{if } s(x) < \xi \end{cases} . \tag{2.1}$$

The particular choice of $s(x)$ depends on the chosen OOD detection method [106]. The primary metrics for evaluating the performance of OOD methods are the Area Under the Receiver Operating Characteristic ($AUROC \uparrow$) curve and the False Positive Rate at 95% True Positive Rate ($FPR95 \downarrow$) [107]. The AUROC is a threshold-independent metric that measures the overall ability of the OOD score $s(x)$ to rank ID samples higher than OOD samples across all possible thresholds. An AUROC score closer to 100% indicates better general separation of the distributions [170]. FPR95, conversely, is a threshold-dependent metric that quantifies the False Positive Rate (FPR), or the rate at which ID samples are incorrectly rejected as OOD, when the threshold ξ is set such that the True Positive Rate (TPR), or the rate at which OOD samples are correctly detected, is at 95%. An FPR95 score closer to 0% is critical for safety-sensitive applications.

We first discuss methods that assume access to labelled ID samples, also referred to as supervised OOD detection. Recent studies have established a connection between ID accuracy and OOD detection [104, 85, 88, 5]. Supervised training produces intermediate representations that form compact, label-related groups [85]. Consequently, many approaches derive anomaly scores directly from ID classifiers. A seminal example is the maximum softmax probability (MSP) score [106], later extended with temperature scaling [170].

Alternative logit-based scores include the maximum logit [103] and the negative energy score [173]. The latter can be refined by trimming the activations of the penultimate layer based on the 90th percentile of activations (global threshold) estimated on the ID training data [252], known as ReAct. NNGuide [204] adjusts the energy score by weighting it with the sample's top-k nearest neighbors' score. Similarly, DICE [253] leverages weight sparsification, retaining the most influential weights for logit computation [251]. ViM [276] fuses class-agnostic features and class-dependent information from the classifier's logits to compute the OOD score.

Another approach is the Mahalanobis distance (MD) based score [158], which assumes class-conditional features follow Gaussian distributions centered at per-class means [85]. Ren et al. adapt the MD score such that a background Gaussian distribution is fitted to the entire training data, disregarding class labels. However, computing the inverse covariance matrix for N training samples, which scales as $O(N^3)$, poses scalability challenges for large datasets. Sun et al. [254] introduced the k-nearest neighbors (k-NN) distance to the ID training set as a simple yet effective OOD detection score. Simple extensions, such as CoVer [307], average the sample's OOD score by applying multiple image perturbations. Pixel-level perturbations include Gaussian noise, weather-like effects (snow, fog), increased contrast or brightness, pixelation, etc [24]. We summarize most of the mentioned OOD detection scores in Table 2.1 to ameliorate the reader.

Method	OOD score $s(x')$, N ID samples, C classes
Using classifier's output (logits) $\upsilon = h(g(x)) \in \mathbb{R}^C$	
MSP [106]	$s_{\text{MSP}}(x') = \max_{c\in\{1,\dots,C\}} \frac{e^{\upsilon'_c}}{\sum_{j=1}^{C} e^{\upsilon'_j}}$
MaxLogit [103]	$s_{\text{MaxLogit}}(x') = \max_{c\in\{1,\dots,C\}} \upsilon'_c$
Energy [173]	$s_{\text{Energy}}(x') = T \cdot \log \sum_{c=1}^{C} e^{\upsilon'_c/T}, \;\; T = 1$
Using intermediate representations $z = g(x) \in \mathbb{R}^d$	
1-NN similarity[254]	$s_{\text{NN}}(x') = \max_i \text{sim}\big(z', z_i^{train}\big)$
Mahalanobis Distance (MD) [158]	$s_{\text{MD}}(x') = -\min_c (z' - \mu_c^{train})^\top \Sigma^{-1} (z' - \mu_c^{train})$, with $\mu_c^{train} = \frac{1}{N_c} \sum_{i:y_i=c} z_i$, and $\Sigma = \Sigma_{train} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} (z_i - \mu_c^{train})(z_i - \mu_c^{train})^\top$
Relative MD (RMD) [221]	$s_{\text{RMD}}(x') = -\min_c \Big[(z' - \mu_c^{train})^\top \Sigma^{-1} (z' - \mu_c^{train}) - (z' - \mu_0)^\top \Sigma_0^{-1} (z' - \mu_0)\Big]$, $\Sigma_0 = \Sigma_{train,0} = \frac{1}{N} \sum_{i}^{N} (z_i - \mu_0)(z_i - \mu_0)^\top$

Table 2.1: **Commonly used OOD detection scores when labeled ID samples are available**. The OOD detection scores that use intermediate representation can also be used without a trained ID classifier with adjustments (k-means cluster assignments can be used in the computation of MD [234]).

Auxiliary data/objectives. Supervised learning alone may not yield sufficiently expressive features for robust OOD detection [285]. To address this limitation, several auxiliary tasks have been proposed to enrich the learned feature representations. Examples include predicting predefined transformations of ID samples [108] or incorporating contrastive objectives. Building on this idea, [194] employs Bayesian optimization to identify ID-specific transformations. Zhang et al. [309] propose a hybrid two-branch framework in which a generative model is trained jointly with a supervised classifier. CIDER [191] is an OOD-specific framework that jointly optimizes a dispersion loss that promotes large angular distances among different class prototypes and a compactness loss that encourages samples to be close to their class prototypes.

External data have also been widely employed [107, 217], often treated as outliers during training with an additional objective, such as entropy maximization. Outlier exposure leads to OOD detection performance gains only when auxiliary data are sufficiently diverse and disjoint from the ID [177]. More commonly, external data are used for pre-training. For example, Fort et al. [85] demonstrate substantial performance gains

when fine-tuning models pretrained on ImageNet-21K.

Synthetic outlier-based methods. Instead of external data, synthetic outliers can be created using generative models [316]. Heng et al. [109] employ unconditional diffusion models trained on ID data to sample likely OOD examples. Kwon et al. [154] segment ID samples and inpaint them with latent diffusion models to produce outliers. VOS [73] generates virtual outliers from the low-likelihood region of the class-conditional Gaussian distribution estimated in the feature space. NPOS [256] extends VOS without making any distributional assumption on the ID embeddings. More recently, Dream-OOD [72] learns a class-conditioned latent space from ID data, then samples low-likelihood latent outliers. These latent outliers can be decoded into images via the trained class-conditional diffusion model. Recently, SynOOD [162] harnesses multi-modal state-of-the-art foundational models to modify ID samples to generate synthetic near-boundary OOD images for fine-tuning CLIP text and image encoder.

OOD detection *without labeled ID samples*

For completeness, we also review unsupervised OOD detection, where only ID samples are available. Unsupervised OOD detection methods typically learn ID-specific features through self-supervised learning [50, 234, 163]. Most unsupervised methods utilize the top-1 nearest neighbor similarity or Mahalanobis distance, based on k-means centroids, as described in Table 2.1. Contrastive-based methods can be further enhanced by designing hand-crafted transformations that provide an estimate of near OOD data or improved OOD scoring functions [255]. For example, Tack et al. [255] add a transformation prediction objective (e.g. rotation prediction such as $[0°, 90°, 180°, 270°]$). Sehwag et al. [234] define a contrastive-based OOD detection method, where the Mahalanobis distance is computed in the feature space using the cluster centroids of k-means. Finally, Rafiee et al. [217] present a self-supervised analog of outlier exposure by rotating auxiliary samples, which are used for entropy maximization along with the self-supervised objective of DINO [42]. Recently, CLIP has enabled learning from vast amounts of raw text [214], leading to a decline in research on purely unsupervised OOD detection.

2.1.3 OOD detection using CLIP

The recent advent of vision-language models (VLMs), exemplified by CLIP [214], has substantially advanced research in computer vision. Beyond the success of CLIP models in general visual understanding, VLMs have had a profound impact on OOD detection [190]. CLIP-based OOD detection has emerged as a rapidly growing and influential area of research [128, 192].

The key assumption of CLIP-based OOD methods relies on the *a priori* knowledge of the finite set of ground truth ID label names $\mathcal{Y}_{\text{GT}} = \{y_1, y_2, \dots\}$ with cardinality $|\mathcal{Y}_{\text{GT}}|$. We refer to the ground truth ID label names as positives. The text encoder encodes these ID label names into representations $\mathcal{Z}_{GT} = [z_1, z_2, \dots z_\nu]^\top \in \mathbb{R}^{|\mathcal{Y}_{\text{GT}}| \times d}$. In general, we denote the representation of any label set $\mathcal{Y}_{labels}$ as $\mathcal{Z}_{labels} = [z_1, z_2, \dots]^\top \in \mathbb{R}^{|\mathcal{Y}_{labels}| \times d}$. Given $\mathcal{Y}_{\text{GT}}$, the classification-based scores in Table 2.1 can be reformulated under a zero-shot inference framework. For a test image x' and its representation $z'_{img} = g_{img}(x') \in \mathbb{R}^d$, the maximum softmax probability is equivalent to the maximum concept matching (MCM)

introduced by Ming et al. [190], where the classification outputs (logits) are image-text similarities $\upsilon' = \mathcal{Z}_{GT} z'_{img} \in \mathbb{R}^{|\mathcal{Y}_{\text{GT}}|}$:

$$s_{\text{MCM}}(x') = \max_{c \in \{1,\dots,|\mathcal{Y}_{\text{GT}}|\}} \frac{\exp(\upsilon'_c)}{\sum_{j=1}^{|\mathcal{Y}_{\text{GT}}|} \exp\big(\upsilon'_j\big)}, \tag{2.2}$$

The energy score is defined similarly. Adaloglou et al. [5] show that pseudo-labels derived from Eq. (2.2) (by computing the arg max instead of max) can be used to train a linear layer, known as probing, on top of the image encoder of CLIP. Combined with the energy score for inference, this approach, termed pseudo-label probing (PLP), achieves strong performance across various CLIP models while remaining computationally lightweight.

2.1.4 Incorporating Negative Concepts into CLIP

By design, CLIP provides a flexible mechanism for specifying which classes are considered ID during inference [88]. CLIP can also leverage a secondary set of unrelated concepts relative to the semantic categories of the ID. We denote the secondary set as $\mathcal{Y}_{\text{neg}}$, representing likely negative (i.e. unrelated) concepts. In general, we denote the primary finite set of ID-related concepts as $\mathcal{Y}_{\text{pos}} = \{y_1, y_2, \dots\}$ that is a proxy of the real semantic categories, such as $\mathcal{Y}_{\text{pos}} = \mathcal{Y}_{\text{GT}}$. The generalized score incorporating both positive and negative concepts, scaled by a temperature $\tau > 0$, is formulated as:

$$\upsilon^{pos} = \mathcal{Z}_{pos} z'_{img} \in \mathbb{R}^{|\mathcal{Y}_{\text{pos}}|}, \quad \upsilon^{neg} = \mathcal{Z}_{neg} z'_{img} \in \mathbb{R}^{|\mathcal{Y}_{\text{neg}}|}, \tag{2.3}$$

$$s_{\text{neg}}(x') = \frac{\sum_{c=1}^{|\mathcal{Y}_{\text{pos}}|} \exp(\upsilon_c^{pos}/\tau)}{\sum_{c=1}^{|\mathcal{Y}_{\text{pos}}|} \exp(\upsilon_c^{pos}/\tau) + \sum_{c=1}^{|\mathcal{Y}_{\text{neg}}|} \exp(\upsilon_c^{neg}/\tau)}. \tag{2.4}$$

The specific choice of $\mathcal{Y}_{\text{pos}}$ and $\mathcal{Y}_{\text{neg}}$ varies across methods. For example, Fort et al. [85] employ Eq. (2.4) with $\mathcal{Y}_{\text{pos}} = \mathcal{Y}_{\text{GT}}$ and $\mathcal{Y}_{\text{neg}} = \mathcal{Y}_{\text{OOD}}$, where $\mathcal{Y}_{\text{OOD}}$ represents the OOD label names. Intuitively, including OOD label names constitutes a weak form of outlier exposure for multi-modal models. However, this approach is not OOD-agnostic and has limited practical applicability. MCM can similarly be extended to incorporate positive and negative concepts (Eq. (2.3)) by replacing the numerator's summation with a max operator:

The specific choice of $\mathcal{Y}_{\text{pos}}$ and $\mathcal{Y}_{\text{neg}}$ varies across methods. For instance, Fort et al. [85] use Eq. (2.4) with $\mathcal{Y}_{\text{pos}} = \mathcal{Y}_{\text{GT}}$ and $\mathcal{Y}_{\text{neg}} = \mathcal{Y}_{\text{OOD}}$ where $\mathcal{Y}_{\text{OOD}}$ represents the OOD label names. Intuitively, including the OOD label names is a weak form of outlier exposure for multi-modal models. However, this approach is not OOD agnostic and finds limited practical applications. MCM can similarly be extended to include positive and negative concepts (Eq. (2.3)) by replacing the nominator's sum with the max operator:

$$s_{\text{neg-mcm}}(x') = \max_{c \in \{1,\dots,|\mathcal{Y}_{\text{GT}}|\}} \frac{\exp(\upsilon_c^{pos}/\tau)}{\sum_{c=1}^{|\mathcal{Y}_{\text{pos}}|} \exp(\upsilon_c^{pos}/\tau) + \sum_{c=1}^{|\mathcal{Y}_{\text{neg}}|} \exp(\upsilon_c^{neg}/\tau)}. \tag{2.5}$$

Incorporating negative concepts for large-scale OOD detection has led to significant progress in FS-OOD benchmarks, demonstrating enhanced robustness to covariate ID shifts. Nonetheless, defining suitable negative concepts for arbitrary ID sets remains challenging. Introducing textual information without access to OOD label names is hence an active area of research using VLMs [297, 192, 180]. For an arbitrary test sample, the ideal negative concept set must be diverse and relevant to the input image, yet non-overlapping with $\mathcal{Y}_{\text{GT}}$ [76]. Current approaches can be divided into three categories: a) training a separate text encoder, b) context prompt learning, and c) mining concepts from a text corpus.

Separate text encoders. ZOC [76] introduces negative text candidates through training an image captioner decoder on paired auxiliary data to generate candidate negative labels per test sample. The OOD score is then computed using Eq. (2.4). However, ZOC can result in overlapping labels with ID data, which is undesirable. To address this, CLIPN [277] trains a dedicated "negative" text encoder to capture negation semantics. An exemplar negative text prompt to create the negative text candidates $\mathcal{Z}_{neg}$ is "a photo of a **no** < label >". Using the notation from Eq. (2.3), the detection score of CLIPN is defined as:

$$s_{CLIPN}(x') = \sum_{c=1}^{|\mathcal{Y}_{\text{GT}}|} (1 - p_c^{neg}) p_c^{pos}, \text{ where} \tag{2.6}$$

$$p_c^{pos} = \frac{\exp(\upsilon_c^{pos})}{\sum_{j=1}^{|\mathcal{Y}_{\text{GT}}|} \exp\big(\upsilon_j^{pos}\big)} \in [0,1], \quad p_c^{neg} = \frac{\exp(\upsilon_c^{neg})}{\exp(\upsilon_c^{pos}) + \exp(\upsilon_c^{neg})} \in [0,1]. \tag{2.7}$$

Here, p_c^{pos} denotes the probability of a sample belonging to class c relative to all other classes, while p_c^{neg} is the probability of class c being a negative instance.

Context prompt learning. Training a separate text encoder requires a large external dataset to improve the quality of negative semantic prompts, while it increases both model complexity and parameter count. To mitigate this, dataset-specific prompts can be learned [318]. More precisely, CLIP employs manually designed prompt templates (context prompt) by default. The different text prompts are averaged for each text candidate. However, manually tuning these context words—known as prompt engineering—is inefficient despite often yielding substantial performance gains [214].

Instead of prompt engineering, context optimization (CoOp) [318] sets some context words in the prompt template as continuous learnable parameters, optimized using few-shot ID samples. CoOp can be adjusted to OOD detection by further regularization. In [193], the authors leverage local features of CLIP, which likely correspond to ID-irrelevant regions, for entropy regularization during CoOp [193]. Similarly, NegPrompt [166] first utilizes CoOp to learn positive prompts and subsequently learns negative prompts for each ID class. Both approaches adopt Eq. (2.5) during inference.

Mining concepts from text corpora. While negative prompt learning can be effective, it typically requires training with multiple objectives. To exemplify, NegPrompt [166] introduces three loss terms that must be balanced during training and has only been

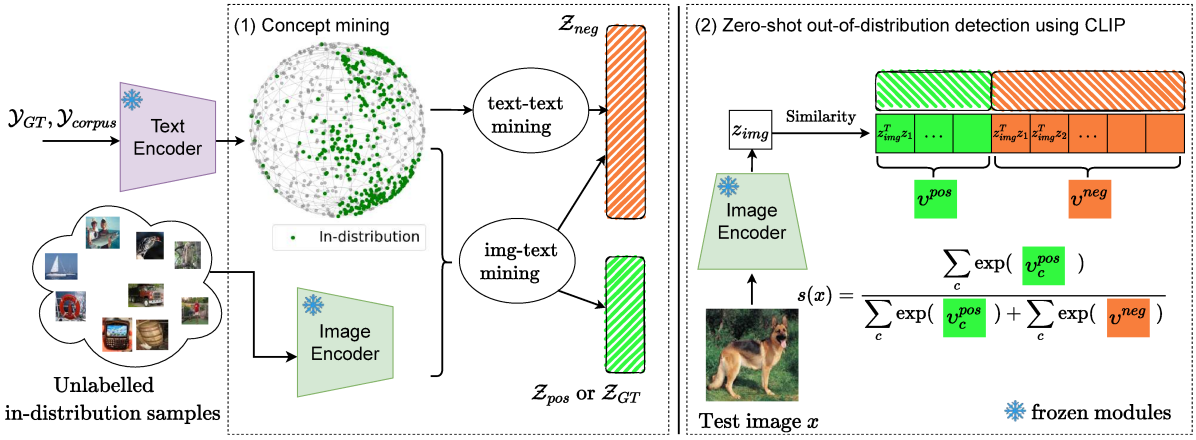


Figure 2.2: **A high-level overview of the concept framework from text corpora for OOD detection using CLIP**. Given $\mathcal{Y}_{\text{GT}}$, negative label mining approaches aim to extract ID-unrelated class names $\mathcal{Y}_{\text{neg}}$ given $\mathcal{Y}_{\text{GT}}$, using text-to-text similarity [131]. $\mathcal{Z}_{\text{neg}}$ is typically realized the most dissimilar text representations from $\mathcal{Z}_{\text{pos}}$. On the other hand, positive label mining methods utilize the shared vision-language space of CLIP first to determine $\mathcal{Y}_{\text{pos}}$. $\mathcal{Z}_{\text{neg}}$ is realized as the non-overlapping elements of $\mathcal{Y}_{\text{pos}}$ and $\mathcal{Y}_{\text{corpus}}$, or using the most dissimilar text representations, as before. At inference time, the OOD detection score is $s(x)$. Best viewed in color.

applied to the smallest publicly available CLIP model, ViT-B, due to computational constraints. A simpler, training-free strategy is to extract concepts directly from a large text corpus using CLIP. An overview of this process is shown in Fig. 2.2.

We now demonstrate how positive and negative concepts can be extracted from a text corpus, beginning with text-to-text similarities. Let the finite text corpus be $\mathcal{Y}_{\text{corpus}} = \{y_1, y_2, \dots\}$ with cardinality $|\mathcal{Y}_{\text{corpus}}|$. Instead of assuming access to OOD label names ($\mathcal{Y}_{\text{neg}} = \mathcal{Y}_{\text{OOD}}$) as in [85], one can mine non-ID concepts. We begin with negative mining, assuming access to $\mathcal{Y}_{\text{pos}} = \mathcal{Y}_{\text{GT}}$. NegLabel [131] extracts $\mathcal{Y}_{\text{neg}}$ using text-to-text similarities. After determining $\mathcal{Y}_{\text{neg}}$, Eq. (2.4) is applied to classify test samples.

Negative label mining (NegMine) computes the negative cosine feature similarity for each $y_{GT} \in \mathcal{Y}_{\text{GT}}$ within $\mathcal{Y}_{\text{corpus}}$. All text encodings are assumed to be L2 normalized. For each concept in the corpus, the distance d_i, where $i = 1, \dots, |\mathcal{Y}_{\text{corpus}}|$ corresponding to the 5th percentile is taken. Instead of the minimum distance, the lower 5th percentile improves robustness against possible outlier ID labels. Finally, K concepts from the corpus with the largest distance are selected.

The primary limitation of NegMine lies in the challenging selection of the hyperparameter K. By splitting $\mathcal{Y}_{\text{neg}}$ into randomly sampled groups, Eq. (2.4) can be applied per negative group [131], which introduces another hyperparameter. The final score is then computed as the average across groups. In general, no consistent performance gains were observed from using this strategy in subsequent works [3]. To maximize coverage of OOD samples, Fu et al. [86] further incorporate label names that are closest to the ground truth labels in the feature space of CLIP, as measured by the smallest percentile distance.

A recent follow-up approach is adaptive negative proxies (AdaNeg [315]). AdaNeg dynamically adjusts the selection of negative labels within the OOD dataset as more

samples become available, aligning the negative labels with the observed OOD samples. Similarly, CLIPscope [86] score integrates the samples' maximum softmax probability computed with $\mathcal{Y}_{\text{pos}}$ in Eq. (2.4) along with an online posterior update, adjusted at inference time. For each test sample x, the posterior update estimates the historical frequency (p_0) of the ID class with the highest image-text similarity to the sample. The historical frequency is an empirical estimation of the likelihood, and the final score is computed as $\frac{1}{p_0} s_{MCM}(x) \cdot s_{neg}(x)$.

Still, the above methods exhibit critical limitations: 1) they assume access to $\mathcal{Y}_{\text{GT}}$, 2) they depend on the choice of negative concepts, and 3) the adapted OOD scores are not OOD-agnostic. When anomalous samples originate from heterogeneous distributions, detection performance tends to degrade. To address these challenges, we develop ClusterMine [3], detailed in the second part of this manuscript. ClusterMine is the first OOD detection method using CLIP that does not assume access to $\mathcal{Y}_{\text{GT}}$, while achieving state-of-the-art OOD detection performance. Fixed ground truth label names may be unavailable, unreliable at scale, or become less relevant due to in-distribution shifts after deployment. Towards truly unsupervised OOD detection, ClusterMine leverages widely available text corpora for positive label mining using image-text similarities, as illustrated in Fig. 2.2. Specifically, ClusterMine [3] extracts positive concepts from a large text corpus by combining visual-only sample consistency (via clustering) and zero-shot image-text consistency. Crucially, using state-of-the-art feature-based clustering with a sufficiently high number of clusters leads to avoiding careful hyperparameter tuning. The negative concepts $\mathcal{Y}_{\text{neg}}$ are defined as the remaining corpus concepts that are not selected as positives, thereby avoiding threshold tuning or adaptive updates to achieve state-of-the-art performance.

2.2 Large-Scale Image Clustering

2.2.1 Introduction

Even when data are noisy or semi-automatically annotated, supervised pre-training on billion-scale image data yields unprecedented robustness [60, 90]. However, what happens when there is no access to paired annotated data to begin with? Clustering is the process of grouping similar samples into distinct groups. More precisely, clustering aims to partition a set of unlabelled data into subsets (the clusters), so that the samples belonging to each subset share common characteristics [300]. Clustering can reveal useful structures in a collection of unlabeled data. For instance, clustering can be used for filtering samples for large-scale pre-training [202] or anomaly detection [3]. Other applications of clustering include generalized category discovery [270] and conditioning generative models [120].

Various clustering algorithms differ in their notion of cluster and their assumptions [79, 300]. Established notions of clusters include: i) samples having higher intra-cluster similarity compared to samples belonging to other clusters [183], ii) connected dense regions in the data space (DBSCAN [78]), and iii) modeling clusters using statistical distributions [188]. In the first case, similarity is measured according to a criterion, such as the Euclidean distance. Additionally, clustering algorithms can be broadly categorized into hierarchical and partitional types. Hierarchical algorithms identify successive clusters using previously established clusters, whereas partitional algorithms determine all clusters simultaneously. Another common division is between parametric and non-parametric clustering algorithms, where parametric refers to specifying the number of clusters C in advance. Another distinction is hard versus soft clustering. Hard clustering refers to each sample being exclusively associated with a single cluster (i.e. k-means, DBSCAN). In contrast, soft clustering allows points to be simultaneously associated with multiple clusters (fuzzy k-means, Gaussian mixture model).

Notably, clustering algorithms are distinguished based on their applicability in large datasets. Here, we use ImageNet-1K, which comprises of 1.2 million samples, as a reference for a sufficiently large dataset, as our focus is on image clustering. Image clustering is the task of assigning a semantic label to an image, given an *a priori* finite set of classes [6]. Ideally, samples within each group perfectly align with the ground truth label. Below, we discuss some limitations of large-scale image clustering:

1. **Identifying the number of clusters is difficult.** Existing methods, such as the elbow method [267], the silhouette score [225], or the minimum coding rate criterion [54], do not consistently approximate the number of ground truth (GT) labels (C_{GT}). The silhouette score measures how well-separated clusters are. For this reason, most parametric approaches consider that C_{GT} is known. Clustering with $C = C_{GT}$ is recently coined as "unsupervised transfer" [87].

2. **Clustering metrics are typically not comparable across different numbers of clusters.** For fair evaluation, it is common to set the number of clusters (C) equal to the ground truth when computing scores such as the normalized mutual information (NMI). When $C = C_{GT}$, one can also compute the clustering accuracy (ACC) using the cluster-class mapping, known as the Hungarian algorithm [153].

3. **Evaluating clustering using ground truth annotations is sub-optimal.**

Human annotators label images using prior world knowledge, are subject to the label collection procedure, and can have varying hierarchies [2, 116, 30]. These biases can lead to systematic inaccuracies when evaluating a clustering algorithm based on a held-out test set of annotated samples. Humans may create visually indistinguishable groups based on arbitrary class distinctions (e.g. the class "laptop computer" and "notebook computer" coexist in ImageNet-1K) [24]. Since clustering algorithms only access visual information, they cannot capture all class granularities, such as the class carnivoran mammals in the coarse annotation level of CIFAR100 [151, 2].

4. **Given C_{GT}, the underlying distribution among classes is tough to learn.** A uniform distribution is often assumed, and benchmarks are created with samples that are uniformly distributed among the GT labels, such as ImageNet-1K and CIFAR100. We refer to these benchmarks as balanced classification benchmarks.

5. **Very few methods report results on the large-scale ImageNet-1K benchmark to date** [268, 12]. It is currently unclear whether methods developed on small datasets can be applied to large-scale benchmarks. This can be due to hyperparameter sensitivity or computational complexity. Spectral clustering is one example with $O(N^3)$ time complexity (N samples). Since we are interested in large-scale image clustering, we will focus on methods that report competitive results on ImageNet-1K.

6. **Real-world images frequently contain multiple objects of interest.** These concepts are not captured when considering hard clustering methods.

Representation learning was a pivotal shift for image clustering [47]. Self-supervised methods, such as contrastive learning and joint-embedding architectures, are designed to be invariant to strong image transformations that preserve label-related information [50, 305, 96]. This is desirable for clustering images, as same-class samples should ideally be grouped together. Representation learning can be viewed as the first step for image clustering, where representations are more separable in the feature space w.r.t. their labels. This can be quantified by the k-NN accuracy. As such, clustering methods that use pretrained encoders (multi-stage clustering) significantly outperform methods that simultaneously aim to learn representations and cluster assignments [301, 12] (single-stage methods). Cluster assignments refer to mapping a sample from image or feature space to a specific cluster index. Apart from the Hungarian mapping, one can associate a cluster with a particular concept by visualizing a few samples or using zero-shot inference using CLIP [53].

Single-stage methods [291, 301, 12] often alternate between learning features and cluster assignments. This is similar to an expectation-maximization process, where k-means is often used for cluster assignments, serving as pseudo-labels for feature learning. For instance, DAC [47] formulates a binary pairwise classification task in which sample pairs are selected based on feature similarity at each iteration, and the resulting pseudo-labels are used to train a convolutional neural network. Similarly, DeepCluster [40] alternates between clustering features via k-means [178] and using the cluster assignments as pseudo-labels to update the model parameters. SeLa [301] further demonstrates that DeepCluster is susceptible to degenerate solutions that can be mitigated through

specific hyperparameter choices. To address this, SeLa introduces a multi-step pseudo-label extraction framework that iteratively estimates the assignment matrix under an equipartition constraint. In PCL [164], clustering is formulated as jointly learning cluster centroids via k-means while optimizing the network through contrastive learning [50]. Building on this, ProPos [122] introduces a proximal contrastive learning framework that mitigates class collisions by maximizing distances between cluster centroids and aligning nearest neighbors [96].

Other end-to-end single-stage approaches aim to learn features and clusters jointly. Such methods are often sensitive to weight initialization [58] and require extensive hyper-parameter tuning, making them computationally expensive at large scales. For instance, DCMM [287] progressively mines NNs and high-confidence samples in feature space, while IIC [129] proposes a mutual information-based objective for paired data. Nonetheless, these approaches rely on stochastic transformations of the same image, mimicking the general-purpose self-supervised methods [74]. More recently, SSCN [12] introduced a single-stage end-to-end method that employs a cross-entropy variant. Although conceptually sound, single-stage frameworks are computationally demanding and have been overshadowed by advances in visual representation learning [50, 42]. To the best of our knowledge, no method has exceeded the 53.5% clustering ACC on ImageNet-1K [212], while naively applying k-means using a state-of-the-art encoder, such as DINO, yields 52.3% ACC. For this reason, we cover multi-stage methods in greater depth below.

2.2.2 Multi-stage Image Clustering Methods

Multi-stage methods decouple representation learning from cluster assignment, benefiting from the advancement in visual self-supervised learning. The advantage of adopting pre-trained encoders becomes more pronounced as encoders are pretrained on larger datasets than the unlabelled images at hand. A notable breakthrough was developed by SCAN using contrastive learned encoders [50, 101]. In more detail, SCAN [268] separates the learning process into three steps: (i) pretraining a network $g(.)$ via contrastive learning, (ii) training a clustering head $\phi(.)$ with parameters θ_ϕ to group nearest neighbors (NNs), and (iii) fine-tuning the encoder using confidently assigned pseudo-labels. The NNs are mined in the feature space once, which can be precomputed. The fine-tuning stage was later shown not to be necessary when adopting the latest state-of-the-art pretrained encoders [6]. Using two samples x, x' from the mini-batch $B = \{(x_1, x'_1), \ldots, (x_{|B|}, x'_{|B|})\}$, where x' is sampled from the NN set in the feature space of g and the output of the head is $q(c|x) = \text{softmax}(\phi(g(x))) \in \mathbb{R}^C$, the semantic clustering loss of SCAN can be formulated as:

$$\mathcal{L}_{\text{SCAN}}(x, x') \coloneqq -\log \underbrace{\sum_{c=1}^{C} q(c|x)q(c|x')}_{alignment} - \alpha \underbrace{\sum_{c=1}^{C} \bar{q}(c) \log \bar{q}(c)}_{-H(\bar{q}(c))}, \quad \alpha > 0, \tag{2.8}$$

where $\bar{q}(c) = \frac{1}{|B|} \sum_{x \in B} q(c|x)$ is the current mini-batch estimate of the cluster distribution. The first term encourages x, x' to be assigned to the same cluster [319]. The second part is an entropy regularization term and aids in avoiding the trivial solution, where all samples are mapped to a single cluster. This is a similar issue with self-supervised repre-

sentation learning. To avoid a bad initialization of the parameters of θ_ϕ, multiple heads κ are usually used, and Eq. (2.8) is averaged for all heads. The final training loss is the average over all heads and all pairs (x, x') in the mini-batch. TSP [319] uses the k-means cluster centers as a weight initialization of the weights of the linear layer to improve training stability and performance $q(x|c) = \text{softmax}(g(x)W_{centroid}),\ W_{centroid} \in \mathbb{R}^{d\times C}$, with d being the feature dimension.

Feature-based clustering using self-distillation. Unlike SCAN, Adaloglou et al. [6] rely on the fact that the encoder is not trained and discard image augmentations as they do not improve performance. This enables precomputing representations and NN, which greatly accelerates training time and convergence. Additionally, a self-distillation framework is introduced where the student head $q_s(c|x) = \text{softmax}(\phi_s(g(x)))$ is trained with backpropagation while the teacher $q_t(c|x) = \text{softmax}(\phi_t(g(x)))$ is optimized using EMA:

$$\mathcal{L}^i_{\text{SCAN}}(x, x') \coloneqq -\log \sum_{c=1}^{C} q_s^i(c|x) q_t^i(c|x') - \alpha \sum_{c=1}^{C} \bar{q_s}^i(c) \log \bar{q_s}^i(c), i \in \{1, \dots, \kappa\}. \tag{2.9}$$

An alternative objective to SCAN, which we present extensively in the second part of the manuscript, is the novel objective called TEMI, proposed by [6]:

$$\mathcal{L}^i_{\text{TEMI}}(x, x') \coloneqq \underbrace{-\frac{1}{\kappa} \sum_{j=1}^{\kappa} \sum_{c'=1}^{C} q_t^j(c'|x) q_t^j(c'|x')}_{\text{instance weighting}} \log \left(\sum_{c=1}^{C} \frac{\left(q_s^i(c|x) q_t^i(c|x')\right)^\beta}{\tilde{q}_t^i(c)} \right), \beta \in (0.5, 1]. \tag{2.10}$$

In Eq. (2.10), $\tilde{q_t}^i(c)$ is an estimate of the teacher cluster distribution. It is computed as an exponential moving average over mini-batches B defined as $\tilde{q_t}^i(c) \leftarrow \lambda\, \tilde{q_t}(c) + (1 - \lambda)\frac{1}{|B|} \sum_{x\in B} q_t^i(c|x)$ for $\lambda \in (0, 1)$. This results in the TEMI objective being less sensitive to the mini-batch size compared to SCAN, which estimates $\bar{q_t}^i(c)$ from the mini-batch. To understand the role of β we examine Eq. (2.10), which consists of two parts inside the log sum: the numerator $\left(q_s^i(c|x) q_t^i(c|x')\right)^\beta$ encourages the class assignment of a positive pair to align and is maximal when this assignment is one-hot. The denominator $\tilde{q}_t^i(c)$ promotes a uniform distribution over clusters. When a cluster has a low probability, it will be up-weighted. Intuitively, β balances these two effects by reducing the influence of the numerator, thus avoiding degenerative solutions. Another advantage is that the hyperparameter β is bounded, and a value of 0.6 consistently works across encoders and downstream datasets.

Still, NNs are noisy, meaning that not all pairs share the same semantic label. To address this, the instance weighting term in Eq. (2.10) given by $w_i(x, x') = \sum_{c=1}^{C} q_t^i(c|x) q_t^i(c|x')$, aims to assigns a higher weight to true positive pairs compared to false positive ones, by combining the outputs of all teacher heads. The losses are symmetrized $\mathcal{L}^i = \frac{1}{2}(\mathcal{L}^i(x, x') + \mathcal{L}^i(x, x'))$ for each head indexed by i. The head with the lowest training loss is used at inference time, and the cluster assignment is computed as $c^* = \arg\max_c q_t(c|x)$.

Feature-based clustering without self-distillation. SCP [167] is another simple approach that trains a clustering head without the need to compute the NN in feature

space. Unlike TEMI [6], it only uses stochastic blurring and random cropping with a 50% probability to construct a pair (x, x'), without self-distillation. The full objective of SCP is formulated as:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{|B|}\sum_{c=1}^{C} q(c|x_i)\log q(c|x_i') \tag{2.11}$$

$$\mathcal{L}_{\text{conf}} = -\log\left(\sum_{i=1}^{|B|}\sum_{c=1}^{C} q(c|x_i)q(c|x_i')\right), \tag{2.12}$$

$$\mathcal{L}_{\text{entropy}} = \sum_{c=1}^{C}\bar{q}(c)\log\bar{q}(c), \qquad \bar{q}(c) = \frac{1}{|B|}\sum_{i=1}^{|B|}(q(c|x_i) + q(c|x_i')), \tag{2.13}$$

$$\mathcal{L}_{SCP} = \mathcal{L}_{ce} + \lambda_1\mathcal{L}_{\text{conf}} + \lambda_2\mathcal{L}_{\text{entropy}}, \text{ where } \lambda_1, \lambda_2 > 0. \tag{2.14}$$

The first term $\mathcal{L}_{ce}$ aligns soft assignments. The second term $\mathcal{L}_{\text{conf}}$ enhances the confidence of cluster predictions, approaching one-hot vectors. Finally, $\mathcal{L}_{\text{entropy}}$ is identical to the entropy regularization term of SCAN using both samples x, x'.

Another approach is to incorporate textual information using CLIP. For instance, TAC [168] enhances image clustering by utilizing the textual semantics of CLIP, using an external corpus (WordNet). Specifically, TAC first selects text concepts that confidently distinguish the images. Subsequently, it jointly trains the text and image modalities of CLIP by mutually distilling cross-modal neighborhood information.

Similar to context optimization in OOD detection [318], pretrained encoders have been adapted for image clustering. This is particularly useful for reconstruction-based methods, such as MAE. The contrastive objective is called NN alignment (NNA) [10] and achieves competitive results after 15-30 epochs on ImageNet-1K. Formally, given a batch of features $Z = [z_1, z_2 \ldots,]^\top \in \mathbb{R}^{|B|\times d}$ and a first-in-first-out queue $\mathcal{Q}$ the NNA objective can be implemented as:

$$\mathcal{L}_{\text{NNA}}(x_i) = -\log\frac{\exp(\text{NN}(z_i, \mathcal{Q})\cdot z_i/\tau)}{\exp(\text{NN}(z_i, \mathcal{Q})\cdot z_i/\tau) + \sum_{j=1}^{|B|}\mathbb{1}_{[j\neq i]}\exp(\text{stop-grad}(z_j)\cdot z_i/\tau)}, \tag{2.15}$$

$$\text{where } \text{NN}(z_i, \mathcal{Q}) = \underset{q\in\mathcal{Q}}{\text{argmax}}(z_i \cdot q), z_i = g(x_i), \tau > 0. \tag{2.16}$$

The adapted weights of $g(.)$ are subsequently used for feature-based clustering.

Current state-of-the-art. TURTLE [87] is the first clustering approach that i) does not rely on image augmentations, ii) does not require precomputing the NN for each sample, iii) employs a 2-step alternating optimization algorithm, iv) is designed to support multiple representation spaces. In the first part of the optimization, (*inner optimization* performed for M_{inner} iterations in each step), an unregularized logistic regression is trained using the gradient descent method. This outputs predictions that are biased towards the direction of the max-margin hyperplane (i.e., SVM [102]). Logistic regression refers to a linear layer that maps features to the clusters. In the second optimization (*outper*

optimization performed for M_{outer} iterations in each step), a clustering head per encoder is trained to match the output of the logistic regression. The clustering head is a linear layer followed by softmax and it is regularized by entropy regularization as in SCAN [268].

Formally, given a mini-batch $B = \{x_1, \ldots, x_{|B|}\}$ and K pretrained visual encoders $g_k(.), k \in [1, \ldots, K]$, let $W_k \in \mathbb{R}^{C \times d_k}, b_k \in \mathbb{R}^C$ be the linear projection parameters of the k-th encoder. In the inner optimization, this linear projection is trained to match the cluster predictions of $q(c|x)$. Importantly, the cluster predictions $q(c|x)$ are the same for all visual encoders, defined as the average prediction of each cluster head $q_k(c|x)$. The inner optimization, performed for M_{inner} iterations, can be formulated as follows:

$$q(c|x) = \frac{1}{K} \sum_{k=1}^{K} q_k(c|x), \tag{2.17}$$

$$\text{where } q_k(c|x) = \text{softmax}(\underbrace{\hat{W}_k g_k(x) + \hat{b}_k}_{\text{not optimized}}), \hat{W}_k \in \mathbb{R}^{C \times d_j}, \hat{b}_k \in \mathbb{R}^C \tag{2.18}$$

$$\mathcal{L}_{\text{inner}} = \sum_{k=1}^{K} \sum_{x \in B} \mathcal{L}_{\text{ce}}(\ q(c|x)\ ,\ W_k g_k(x) + b_k), \tag{2.19}$$

$$\mathcal{L}_{\text{ce}}(p, q) = -\sum_{i=1}^{C} p_i \log q_i. \tag{2.20}$$

In the inner optimization, all W_k, b_k are trained without regularization nor softmax, optimized via gradient descent. For each visual encoder, $\hat{W}_k, \hat{b}_k$ are randomly initialized before the first inner optimization. The bracket indicates that no gradients are computed, known as stop gradient. The outer optimization, performed for M_{outer} iterations, optimizes $\hat{W}_k, \hat{b}_k$ through the predictions $q(c|x)$, computed as stated above. Finally, the outer optimization objective is formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{outer}} &= \sum_{k=1}^{K} \sum_{x \in B} \mathcal{L}_{\text{ce}}(q(c|x)\ ,\ \underbrace{W_k h_k(x) + b_j}_{\text{not optimized}}) + \lambda \mathcal{L}_{\text{entropy}}, \lambda > 0 \\ \mathcal{L}_{\text{entropy}} &= \sum_{k=1}^{K} \sum_{c=1}^{C} \bar{q}_k(c) \log \bar{q}_k(c), \qquad \bar{q}(c) = \frac{1}{|B|} \sum_{x \in B} q(c|x), \end{aligned} \tag{2.21}$$

After training, cluster assignments are computed as $c^* = \arg\max_c q(c|x)$ for each sample.

2.3 Image Synthesis with Diffusion Models

2.3.1 Introduction

Visual generative modeling has seen rapid progress over the past decade. Multiple paradigms have been proposed for image synthesis. Early approaches, such as variational autoencoders (VAEs) [145], model the data distribution through latent representations, enabling efficient sampling but often producing blurry outputs. Generative adversarial networks (GANs) [92] use an adversarial training scheme between a generator and discriminator, achieving highly realistic image synthesis yet suffering from training instability and mode collapse. More recently, flow-based models [69, 146] have offered precise likelihood estimation using invertible transformations.

Among existing state-of-the-art approaches, diffusion models [244, 113] have emerged as a powerful alternative that combines relatively stable training with high perceptual quality. More importantly, diffusion models offer diverse sampling capabilities, allowing additional control of the generated outcome [313]. There are various perspectives in the literature that have led to similar learning objectives. We present a timeline of seminal works in Fig. 2.3. Even though these perspectives may seem different at first sight, they are connected [77, 156]. They all describe the flow of probability density starting from a simple distribution (e.g. Gaussian) and moving towards the data distribution. For instance, the variational perspective originates from Variational Autoencoders (VAEs) [145], which we describe below.

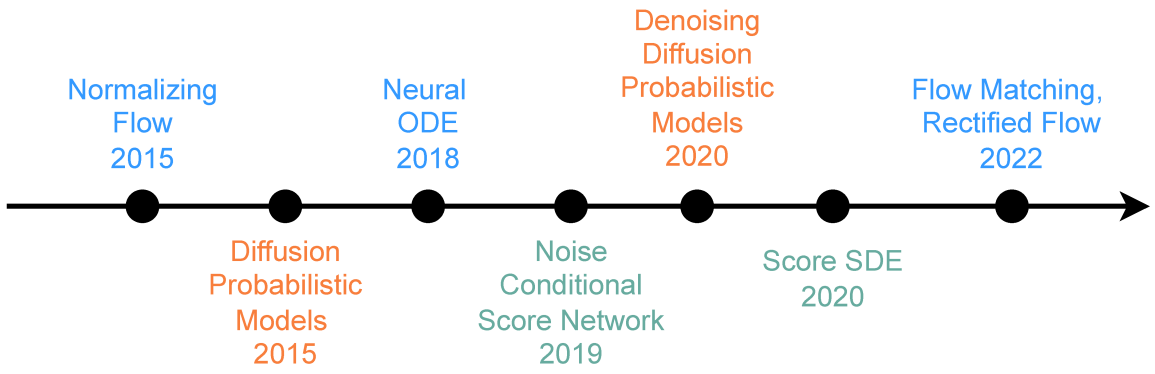


Figure 2.3: **Timeline of diffusion model perspectives.** Each group (marked by different colors) highlights a different perspective on diffusion models. We mainly present the perspective of Variational Autoencoder (VAE) [144] → Diffusion Probabilistic Models (DPM) [244] → Denoising Diffusion Probabilistic Models (DDPM) [113]. Some connection with the group Noise Conditional Score Network [248] → Score SDE [249]. The perspective of the group: Normalizing Flow (NF) [223] → Neural ODE (NODE) [49] → Flow Matching (FM) [172], Rectified Flow [174] is not covered in this work.

2.3.2 Diffusion models: forward and backward process

Diffusion models (DMs) [244, 113] are a class of generative models that learn to reverse a noising process. Here, we present the perspectives of understanding DMs through two Markov processes: the forward process and the learned backward process [203]. A

Markov process or fixed Markov chain means that each step only depends on the previous one. The *forward (noising) process* progressively perturbs real data with Gaussian noise, and the *reverse (denoising) process*, which learns to reconstruct data from noise. DMs synthesize images by learning to reverse a gradual noising process.

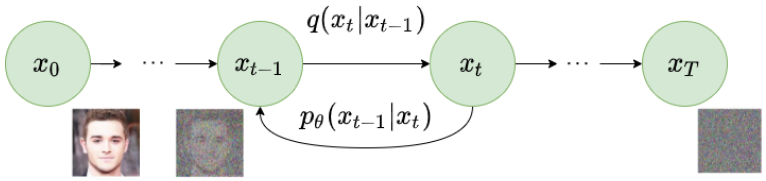


Figure 2.4: **A visual illustration of the forward diffusionand backward/reverse diffusion process.**

Forward process

In DMs, the forward process iteratively adds Gaussian noise (Fig. 2.4). Let $\mathbf{x} \sim q(\mathbf{x})$ denote a real image sample drawn from the data distribution $q(\mathbf{x})$. From now on, we use $\mathbf{x}_0$ to denote the clean sample. The forward diffusion process is defined as a fixed Markov chain that gradually corrupts $\mathbf{x}_0$ over T discrete timesteps, where T controls the total diffusion duration. Specifically, at each timestep $t \in \{1, \ldots, T\}$, Gaussian noise is added according to

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t|\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\big), \tag{2.22}$$

where $\beta_t \in (0, 1)$ is a pre-defined variance schedule that determines the noise magnitude at step t. $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu} = \sqrt{1-\beta_t}\mathbf{x}_{t-1}$ and covariance $\boldsymbol{\Sigma} = \beta_t\mathbf{I}$. Since the covariance matrix is diagonal with shared variance β_t, the noise is sampled independently for each pixel. We can express the distribution of the noisy sample $\mathbf{x}_t$ at any timestep t, based solely on $\mathbf{x}_0$ [244, 247]

$$\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}\big(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{2.23}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{r=1}^{t} \alpha_r$. Eq. (2.23) can be rewritten using the reparameterization trick as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2.24}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are samples drawn from the isotropic Gaussian distribution. Here, $\mathbf{x}_t$ is a linear combination of the clean sample $\mathbf{x}_0$ and independent Gaussian noise scaled by $\sqrt{1-\bar{\alpha}_t}$. Intuitively, $\sqrt{\bar{\alpha}_t}$ controls the signal strength and $\sqrt{1-\bar{\alpha}_t}$ determines the noise level at each timestep. Eq. (2.24) follows from the *reparameterization trick*, a technique that isolates the stochasticity in $\boldsymbol{\epsilon}$, while keeping the transformation with respect to $\mathbf{x}_0$ and model parameters differentiable.

Variance preservation across timesteps is an intentional property of Eq. (2.24) [244, 113]. Specifically, since both $\mathrm{Var}[\mathbf{x}_0] = 1$ and $\mathrm{Var}[\boldsymbol{\epsilon}] = 1$, the coefficients $\sqrt{\bar{\alpha}_t}$ and $\sqrt{1-\bar{\alpha}_t}$ are designed such that $\mathrm{Var}[\mathbf{x}_t] = \bar{\alpha}_t\,\mathrm{Var}[\mathbf{x}_0] + (1-\bar{\alpha}_t)\,\mathrm{Var}[\boldsymbol{\epsilon}] = 1$. This ensures that the forward process maintains a consistent scale of the data while progressively corrupting the signal. Variance exploding formulations are also frequently used [249, 136], i.e. by setting $\sqrt{\bar{\alpha}_t} = 1$.

Backward process

Starting from a noisy sample $\mathbf{x}_T$, the learned backward process gradually denoises samples towards the data distribution $q(\mathbf{x}_0)$ (Fig. 2.4). This is achieved by learning a Markov chain parameterized by a neural network with parameters θ as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}|\boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\big), \quad t \in \{T, \dots, 1\}, \tag{2.25}$$

where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ can be learnable functions. Ideally, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ approximates $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$; however, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is intractable. Sohl-Dickstein et al. [244] demonstrated that it becomes tractable by additionally conditioning on $\mathbf{x}_0$. In other words, we can sample $\mathbf{x}_t$ at noise level t conditioned on $\mathbf{x}_0$. When $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is conditioned on $\mathbf{x}_0$, it can be proved that [181, 113]:

$$\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}), \sigma_q^2(t)\mathbf{I}), \tag{2.26}$$

$$\text{where} \quad \sigma_q^2(t) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{2.27}$$

$$\text{and} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0, \tag{2.28}$$

$$\text{or equivalently} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\Big). \tag{2.29}$$

To arrive in the last equation, we use the parametrization trick from Eq. (2.24) such that $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\ \boldsymbol{\epsilon})$. Eq. (2.29) shows that $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, t)$ can be written only as a function of the noisy sample $\mathbf{x}_t$ and $\boldsymbol{\epsilon}$. A parametrized model $\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t)$ with parameters θ can be trained to predict the mean $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, t)$. Using Eq. (2.28) the parametrized model $\mathbf{x}_\theta(\mathbf{x}_t, t)$ predicts the target $\mathbf{x}_0$. However, it is more common to predict the noise, as shown in Eq. (2.29)), with a parametrized model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ that predicts $\boldsymbol{\epsilon}$. In both cases, $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$. Possible design choices for $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ are discussed below.

2.3.3 Training objective of DMs

Recall that we need to train the neural network that yields $\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t)$ to approximate the mean of $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ in the reverse diffusion process. With the additional conditioning of $\mathbf{x}_0$, it is proved that the evidence lower bound (ELBO) can be formulated as [181]:

$$\log p_\theta(\mathbf{x}_0) \geq \underbrace{\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\big(q(\mathbf{x}_T|\mathbf{x}_0) \,\|\, p(\mathbf{x}_T)\big)}_{\text{prior matching term}}$$

$$- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\big[D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big)\big]}_{\text{denoising matching term}} = -\mathcal{L}_{ELBO}, \tag{2.30}$$

where the reconstruction and prior matching terms are constant w.r.t. the network parameters θ and can be estimated using standard techniques [144]. The proof can be found in [181, 113]. Ignoring the constants, the DM objective that is minimized becomes

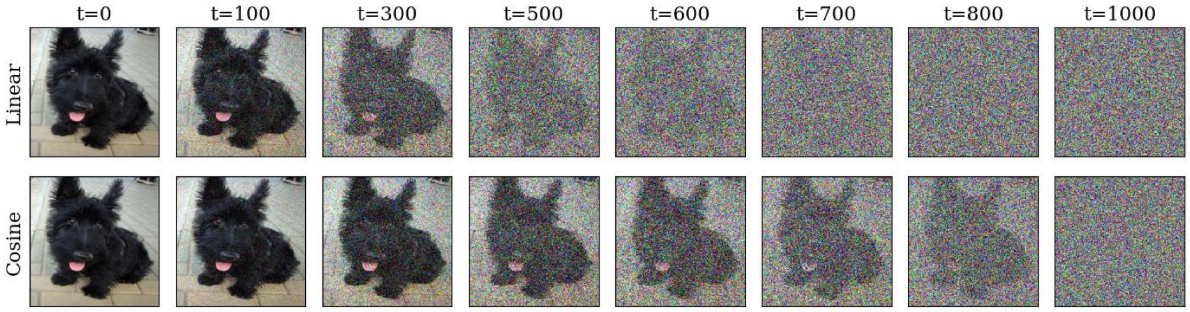


Figure 2.5: **Visualization of the linear versus cosine variance schedule (rows) for β_t across different timesteps, indicated by the columns.**

$$\mathcal{L}_{DM} = \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \big[D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big)\big] \tag{2.31}$$

By setting $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_q^2(t)$, optimizing the KL divergence reduces to minimizing the difference between the target $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, t)$ and parametrized mean $\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t)$. Up to an additive constant, the loss reduces to the squared distance between the means:

$$D_{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) = \frac{1}{2\sigma_q^2(t)} \big\| \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t) \big\|_2^2, \tag{2.32}$$

By substituting the mean expressions from above based on the noise or target parametrization, we can formulate the objective based on noise or target prediction equivalently as:

$$D_{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) = \underbrace{\frac{(1-\alpha_t)^2}{2\ \sigma_q^2(t)\alpha_t(1-\bar{\alpha}_t)}}_{\text{weightning term}} \big\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \big\|_2^2, \tag{2.33}$$

$$\text{or for target prediction} \quad \frac{(1-\alpha_t)^2 \bar{\alpha}_{t-1}}{2\ \sigma_q^2(t)\ (1-\bar{\alpha}_t)^2} \big\| \boldsymbol{x}_0 - \boldsymbol{x}_\theta(\mathbf{x}_t, t) \big\|_2^2 \tag{2.34}$$

The ELBO objective can be approximated using mini-batches and stochastic gradient descent by minimizing Eq. (2.32) using an expectation over timesteps (i.e. using a uniform distribution $t \sim \mathcal{U}(2, T)$).

Choosing $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ and parametrization of noise schedules. To achieve state-of-the-art performance, one needs to carefully design β_t. For instance, Ho et al. [113] employed a linear variance schedule for β_t in $[10^{-4}, 2 \cdot 10^{-2}]$. Oddly, these values are relatively small compared to the normalized image pixels in $[-1, 1]$. As illustrated in Fig. 2.5 (top row), the end of the linear forward noising process is dominated by noise [200], and thus it does not contribute to achieving high perceptual quality. To address this, Nichol et al. [200] use a cosine schedule, as shown in Fig. 2.5 (bottom row).

Parameterizing $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ into the variational bound loss $\mathcal{L}_{DM}$ leads to unstable training [200]. The choice of $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_q^2(t)\mathbf{I}$ in Ho et al. [113] aligns with Eq. (2.34). However,

this choice was experimentally inferior to the unweighted objective in [113], which is partially attributed to the poor choice of the schedule of β. For this reason, Ho et al. [113] employed the unweighted mean squared error (MSE)

$$\mathcal{L}_{MSE}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta) = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)||_2^2 = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}\ ,\ t)||_2^2. \tag{2.35}$$

In this case, the training objective becomes a simple denoising score-matching loss. Later on, Nichol et al. [200] found it better to parameterize the variance $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ as an interpolation between β_t and $\sigma_q^2(t)$ by predicting a vector $\mathbf{v}$ using a neural network, such that

$$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \exp\big(\mathbf{v}\log(\beta_t) + (1-\mathbf{v})\log\big(\sigma_q^2(t)\big)\big), \tag{2.36}$$

$$\mathcal{L}_{IDM} = \mathcal{L}_{MSE}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta) + \lambda\mathcal{L}_{DM}, \tag{2.37}$$

where λ is typically of order 10^{-3}.

Noise weighting and signal-to-noise ratio (SNR). An intuitive prism to study the diffusion objective is the signal-to-noise ratio (SNR). The SNR quantifies the ratio of the clean data component compared to the Gaussian noise at each timestep. Given $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, the SNR is defined as $\text{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}$. As forward diffusion progresses, $\bar{\alpha}_t$ typically decreases monotonically, leading to a decaying SNR. Mathematically, $\text{SNR}(t) < \text{SNR}(t-1)$ for $t \geq 1$. Among others, Luo [181] showed that the weighting term of the training objective is related to the SNR

$$D_{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big) = \frac{1}{2}\underbrace{\big(\text{SNR}(t-1) - \text{SNR}(t)\big)}_{>0}\mathcal{L}_{MSE}(\mathbf{x}_0, \mathbf{x}_\theta) \tag{2.38}$$

Crucially, the SNR plays a key role in designing parametrized noise schedules during training [113, 143]. For instance, we can directly parameterize the SNR at each timestep using a monotonically increasing neural network $\lambda(t)$ with parameters θ_λ, and learn it jointly along with $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ [143]. The log singal-to-noise ratio is defined as $\lambda(t) = -\log(\text{SNR}(t))$. Now, let $T \to \infty$, $dt > 0$, and $\text{SNR}'(t) = d\text{SNR}(t)/dt$. It is shown that the continuous-time diffusion loss simplifies to [143, 77]:

$$L_\infty(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}\{2,T\}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\underbrace{\left(\frac{SNR(t-dt) - SNR(t)}{dt}\right)}_{-\text{SNR}'(t)>0}\mathcal{L}_{MSE}(\mathbf{x}_0, \mathbf{x}_\theta)\right] \tag{2.39}$$

By further defining $\lambda'(t) = d\lambda'(t)/dt = -\frac{\text{SNR}'(t)}{\text{SNR}(t)} = -\text{SNR}'(t)\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$, the continuous-time objective can be formulated in terms of the error prediction as [142]:

$$L_\infty(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}\{2,T\}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\underbrace{\lambda'(t)}_{\text{loss weight}}\mathcal{L}_{MSE}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta)\right], \tag{2.40}$$

where $\mathcal{L}_{MSE}(\boldsymbol{x}_0, \boldsymbol{x}_\theta) = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\mathcal{L}_{MSE}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta)$. In state-of-the-art DMs, a weighting function $\omega(t, \lambda'(t)) > 0$ is employed instead of $\lambda'(t)$ that weights the contribution of different noise levels [136, 230, 172, 99]. The main intuition behind noise weighting is to upweight intermediate timesteps by oversampling them [77]. For a systematic overview of weighting functions, we refer the reader to [142, 156]. Effectively, *differences in diffusion training objectives lie in the design of* $\omega(t, \lambda'_t)$ [156].

2.3.4 Connection with noise-conditioned score-based networks

The connection between the variational DM perspective and other formulations, such as Song & Ermon's [248] score-based generative modeling, reveals similar underlying training objectives. Score-based generative modeling refers to samples produced via Langevin dynamics [282] by using gradients of the data distribution, which are estimated with score matching. The score of $q(\mathbf{x})$ is defined as $s(\mathbf{x}) = \nabla_{\mathbf{x}} \log q(\mathbf{x})$. Langevin dynamics iteratively update a sample $\mathbf{x}$ using this score to draw samples from $q(\mathbf{x})$, and a score model, $\mathbf{s}_\theta(\mathbf{x})$, is used as an approximation of $s(\mathbf{x})$.

Similar to the variation perspective, Song & Ermon [248] perturb the data with Gaussian noise based on a noise schedule. Then, they train the noise-conditioned score network ($\mathbf{s}_\theta(\mathbf{x})$) to estimate the scores. This schedule of gradually increasing noise levels closely resembles the described forward process of DMs. This connection allows the score to be written in terms of the predicted noise $\boldsymbol{\epsilon}_\theta$, given the relation established in Eq. (2.23) and Eq. (2.24), resulting in:

$$\mathbf{s}_\theta(\mathbf{x}_t, t) \approx \mathbb{E}_{q(\mathbf{x})}[\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x})] = \mathbb{E}_{q(\mathbf{x})}\left[-\frac{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}}\right] = -\frac{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}} \tag{2.41}$$

The mathematical equivalence $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}) = -\frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}}$ is derived from the reparameterization of $\mathbf{x}_t$. Hence, learning $\mathbf{s}_\theta(\mathbf{x}_t, t)$ is equivalent to modeling the negative of the source noise $\boldsymbol{\epsilon}$ up to the scaling factor $1/\sqrt{1-\bar{\alpha}_t}$. Intuitively, since the noise is added to the image, moving in the opposite direction "denoises" the image [181]. We refer the reader to [156] for a detailed analysis of score-based methods and their connections to other diffusion-based formulations. In summary, we have established three equivalent formulations for the training objective of DMs, where the neural network is trained to: i) predict the clean image, ii) predict the noise $\boldsymbol{\epsilon}$, or iii) estimate the score $s(\mathbf{x})$ of the image at an arbitrary noise level.

2.3.5 Sampling from diffusion models

Once the score network (or $\boldsymbol{\epsilon}$-prediction network) is trained, the sampling process starts from a pure noise sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoises it back to $\mathbf{x}_0$. This reverse process requires estimating the clean sample $\mathbf{x}_0$ at each timestep t. Using the reparameterized formulation $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ the estimation of the clean sample $\mathbf{x}_0$, denoted as $\hat{\mathbf{x}}_0$, is formulated as [246, 247]

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\Big). \tag{2.42}$$

The full iterative (reverse) step samples $\mathbf{x}_{t-1}$ from the approximated posterior distribution: $\mathbf{x}_{t-1} \sim \mathcal{N}\big(\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t),\ \sigma_q^2(t)\mathbf{I}\big)$. This update can be more transparently expressed by decomposing the mean $\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t)$ into a weighted combination of the predicted clean sample $\hat{\mathbf{x}}_0$ and the current noise estimate $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z} \tag{2.43}$$

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\underbrace{\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}\right)}_{\text{“predicted } \hat{\mathbf{x}}_0\text{”}} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_q^2(t)}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}_{\text{“direction pointing to } \mathbf{x}_t\text{”}} + \sigma_q(t)\boldsymbol{z}, \tag{2.44}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\sigma_q^2(t)$ is the variance of the reverse process, controlling its stochasticity. The first term reconstructs the predicted clean image $\hat{\mathbf{x}}_0$, while the second guides the update toward $\mathbf{x}_t$. The third term introduces the actual stochastic noise required for sampling from the reverse distribution. In practice, a deterministic variant is often used, such as denoising diffusion implicit models (DDIM) [247], which omits the stochastic noise component ($\sigma_q^2(t) = 0$). DDIM is a direct interpolation between $\hat{\mathbf{x}}_0$ and the $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$. Finally, it is common to use a different noise and often shorter noise schedule for sampling than the one used during training [143].

2.3.6 Latent-space diffusion

While the formulation above is defined in pixel space, operating directly on high-dimensional image data is computationally expensive. To improve efficiency, modern approaches such as *Latent Diffusion Models (LDMs)* [224] perform diffusion in the latent space of a pre-trained autoencoder. Let $\mathbf{z} = g(\mathbf{x})$ denote the encoded latent representation. The encoder downscales the spatial dimensions by a factor of 8. However, some structure is maintained in the grid-based latent representation [210].

Diffusion is then applied to $\mathbf{z}$ instead of $\mathbf{x}$, significantly reducing computational cost while maintaining high visual fidelity. After denoising, the decoder converts the latent representation to an image. We use $\mathbf{x}$ to denote latent representations when we adopt LDMs for simplicity.

2.3.7 Conditioning and Guiding DMs at Sampling Time

Thus far, our analysis has focused on approximating the data distribution $p(\mathbf{x})$, known as unconditional generation. Nonetheless, DMs can be conditioned on auxiliary information. When dealing with labeled data, we typically aim to learn a conditional distribution $p(\mathbf{x}|y)$ by predicting $\mathbf{x}_\theta(\mathbf{x}, t, \mathbf{e}_y)$, $\boldsymbol{\epsilon}_\theta(\mathbf{x}, t, \mathbf{e}_y)$, or $\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{e}_y)$. Here, $\mathbf{e}_y$ is the representation of y. For instance, a text representation $\mathbf{e}_y$ derived from a conditioning prompt y (e.g., via CLIP [214]), or a one-hot encoding of the class index.

Conditioning allows us to control the data we generate explicitly, where y is usually combined with the timestep information. For instance, class labels are represented as a one-hot encoding vector and are added to a vector representing the timesteps [136]. In practice, diffusion models can exhibit poor sample quality or ignore the conditioning

Figure 2.6: **Clean samples (top row) and their 3-channel PCA latent visualization (bottom row) using Stable Diffusion-XL (SD-XL) [210].**

signal. To address this, diffusion guidance techniques have been developed to control the impact of the conditioning signal on the generated outcome [137]. Upweighting the condition frequently comes at the cost of deviating from the data distribution [4].

In this section, we use the score-based formulation of a diffusion model $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$. A principled perspective for guidance-based sampling in DMs is the decomposition of the conditional score using Bayes' rule. For each timestep t, we can derive the following equivalent form (recall that $\nabla_{\mathbf{x}_t} \log q_t(y) = 0$):

$$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) = \nabla_{\mathbf{x}_t} \log \left(\frac{q_t(y|\mathbf{x}_t) q_t(\mathbf{x}_t)}{q_t(y)} \right) = \underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)}_{\text{unconditional direction}} + \underbrace{\nabla_{\mathbf{x}_t} \log(q_t(y|\mathbf{x}_t))}_{\text{guidance direction}} \tag{2.45}$$

Instantiations of the Guidance Direction

Classifier Guidance (CG). Nichol et al. [200] train a classifier $f(y|\mathbf{x}_t, t)$ with parameters θ_f to approximate $q_t(y|\mathbf{x}_t)$, which allows for gradients w.r.t. $\mathbf{x}_t$ to be computed at each step. By combining Eq. (2.41) and introducing a scaling factor $(w+1) \geq 0$ for fine-grained control of the strength of the guidance direction, we arrive at the CG sampling equation:

$$\nabla_{\mathbf{x}_t} \log \tilde{q}_t(\mathbf{x}_t|y) \approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \underbrace{\left(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - (w+1)\sqrt{1-\bar{\alpha}_t} \nabla_{\mathbf{x}_t} f(y|\mathbf{x}_t, t) \right)}_{\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)}, \tag{2.46}$$

When $w = -1$, the condition is ignored (unconditional generation). For $w = 0$, we recover the true conditional score. Sufficiently large positive values of $w > 0$ amplify the guided direction. This ensures condition adherence and high perceptual quality in the generated image, often at the expense of diversity (Fig. 2.7). A significant drawback of CG is its reliance on the separately learned noisy classifier, which needs to be trained jointly with the diffusion model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$.

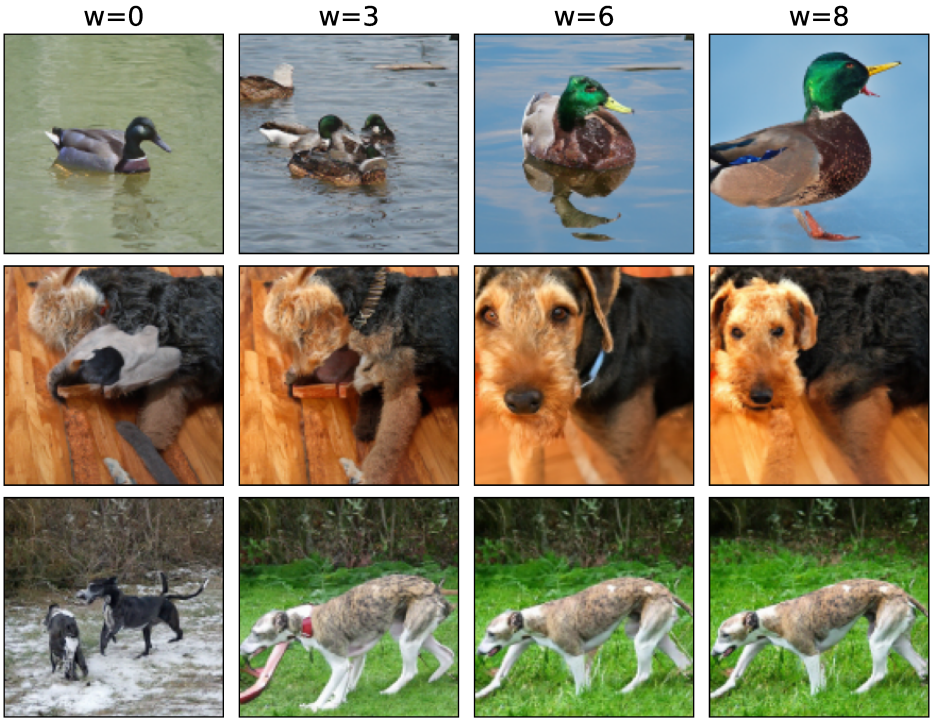


Figure 2.7: **Generated samples guided using classifier guidance (CG) with varying guidance weight w.** The ADM [66] model is used, trained on ImageNet-1K (128×128 resolution).

Classifier-free guidance (CFG). Ho et al. [114] showed that it is possible to guide the conditional generation, *without an independent classifier*, by combining the scores from a conditional and an unconditional DM. We can re-apply the Bayes' rule to $q_t(y|\mathbf{x}_t)$, effectively inverting the classifier to a generative model $\nabla_{\mathbf{x}_t} \log q_t(y|\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$. Substituting the latter in Eq. (2.45) and adding the guidance scale term $(w+1) \geq 0$ leads to:

$$\nabla_{\mathbf{x}_t} \log \tilde{q}_t(\mathbf{x}_t|y) = \underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)}_{\text{unconditional}} + (w+1)\Big(\underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y)}_{\text{conditional}} - \underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)}_{\text{unconditional}}\Big) \tag{2.47}$$

$$\approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \underbrace{\big((w+1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{e}_y) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\big)}_{\tilde{\boldsymbol{\epsilon}}_{\theta,w}(\mathbf{x}_t, t, \mathbf{e}_y)}, \tag{2.48}$$

Ultimately, CFG modulates the noise prediction network as follows:

$$\tilde{\boldsymbol{\epsilon}}_{\theta,w}(\mathbf{x}_t, \mathbf{e}_y) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{e}_y) + w[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{e}_y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{e}_\varnothing)], \tag{2.49}$$

where the timestep t is omitted for simplicity, $w \geq -1$ is the guidance scale, and $\mathbf{e}_\varnothing$ denotes the encoding of an empty string $\varnothing$, or no class index. While mathematically it is possible to apply the Bayes' rule, Ho et al. [114] argue that there is no performance guarantee of such an approach. It is not trivial to assume that inverting a generative model to a classifier and vice versa yields a useful guidance signal [114].

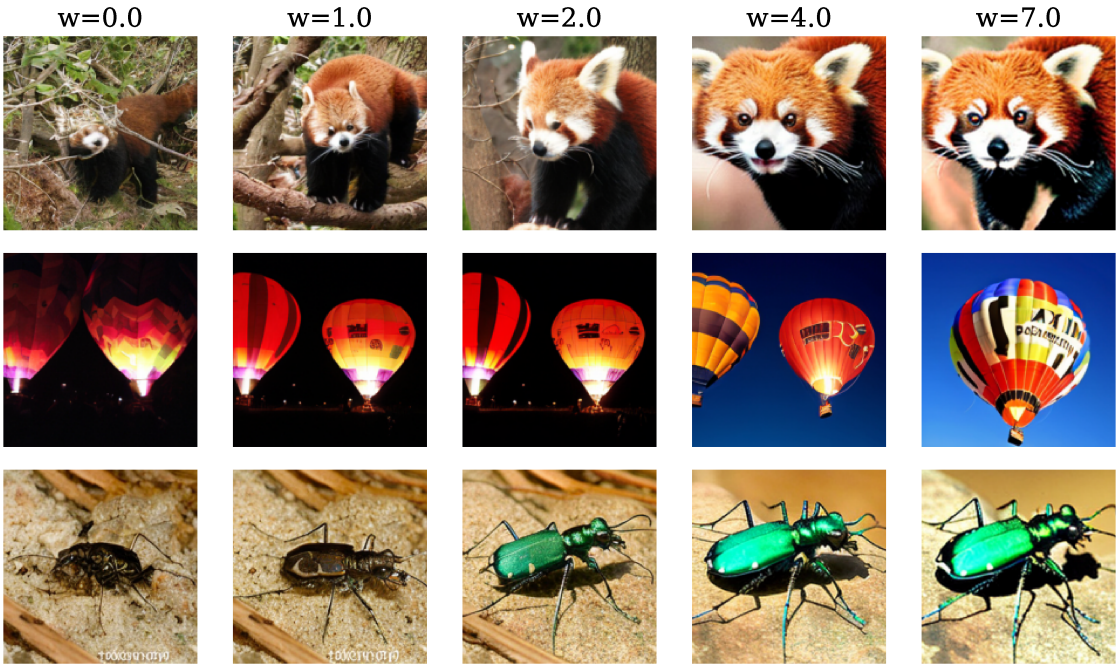


Figure 2.8: **Generated samples guided by CFG at 256×256 resolution with varying guidance weight w using DiT-XL [206] trained on ImageNet-1K**. High guidance leads to excessive saturation and contrast [4].

Empirically, CFG achieves a similar tradeoff in terms of FID [110] and IS [22, 45] as CG. The conditional and unconditional predictors are learned jointly via a single neural network. Precisely, a conditional diffusion model is trained on paired data. During training, conditioning is randomly replaced with $\varnothing$, discarding the condition. This simple training strategy enables a single DM to learn both score functions and linearly combine them at inference time by tuning w. In Fig. 2.8, we illustrate the impact of w for class-conditional generation on ImageNet. In the extrapolation regime ($w > 0$), the DM moves away from the unconditional outcome.

Limitations of CFG and DM guidance variants

CFG exhibits notable limitations. Apparently, it can only be applied to conditional generation (labeled datasets), while trained conditional-only denoisers cannot be used. Moreover, when both denoisers are trained jointly, the sampling trajectory can overestimate the conditional data distribution. This leads to a skewed sample distribution, and oftentimes the generated samples have oversimplified image compositions, as well as an undesirably high contrast and saturation (Fig. 2.8, w=7) [137, 228, 4].

CFG can be improved by gradually adding a noise schedule to the condition [227] or introducing monotonically increasing guidance weight schedules [280]. Saharia et al. [229] apply stepwise intensity thresholding to circumvent overshooting the data distribution. Similarly, Kynkäänniemi et al. apply CFG only at intermediate denoising steps [155]. Sadat et al. [228] decompose the unconditional model into an orthogonal and parallel component and downweight the parallel one, which empirically leads to excessive saturation. Karras et al. train a separate unconditional DM with significantly fewer

parameters [137], while Papalampidi et al. [203] leverage online feedback from a suite of small-scale evaluators to dynamically adjust the guidance scale.

Alternative guidance methods. Recent guidance methods have been developed that deviate from relying on the unconditional model as in CG and CFG. These approaches aim to tackle some of the limitations of CFG by generalizing its linear extrapolation scheme between a positive and negative denoiser:

$$\tilde{\boldsymbol{\epsilon}}_{\theta,w}(\mathbf{x}_t, t, \mathbf{e}_y) = \underbrace{\boldsymbol{\epsilon}_{\text{positive}}(\mathbf{x}_t, t, \mathbf{e}_y)}_{\text{conditional or not}} + w[\boldsymbol{\epsilon}_{\text{positive}}(\mathbf{x}_t, t, \mathbf{e}_y) - \underbrace{\boldsymbol{\epsilon}_{\text{negative}}(\mathbf{x}_t, t, \mathbf{e}_y)}_{\text{inferior performance}}] \tag{2.50}$$

$$= \boldsymbol{\epsilon}_{\text{pos}}(\mathbf{x}_t) + w[\boldsymbol{\epsilon}_{\text{pos}}(\mathbf{x}_t) - \boldsymbol{\epsilon}_{\text{neg}}(\mathbf{x}_t)], \tag{2.51}$$

CFG is a special case of Eq. (2.51), where the negative model is the unconditional denoiser conditioned on $\mathbf{e}_\varnothing$. The rationale of Eq. (2.51) lies in designing a negative model $\boldsymbol{\epsilon}_{\text{neg}}(\mathbf{x}_t)$ that has an inferior performance compared to a well-performing DM, which we call positive. Adaloglou et al. [4] show that guidance works best when "*the negative model makes similar errors to the positive model, but stronger*". Instantiations of $\boldsymbol{\epsilon}_{\text{neg}}(\mathbf{x}_t)$ include:

1. **Reduced training**: $\boldsymbol{\epsilon}_{\text{neg}}(\mathbf{x}_t)$ is a less trained DM compared to $\boldsymbol{\epsilon}_{\text{pos}}(\mathbf{x}_t)$ [137]. Common values are $\frac{1}{8}, \frac{1}{16}$ of the training time of the postive model.

2. **Reduced capacity and/or reduced training**: Reduced capacity trains a secondary diffusion model with fewer parameters compared to $\boldsymbol{\epsilon}_{\text{pos}}(\mathbf{x}_t)$ [137]. Reduced capacity can be combined with reduced training, known as "AutoGuidance" [137].

3. **Finetuning-based methods** include additional training of $\boldsymbol{\epsilon}_{\text{pos}}(\mathbf{x}_t)$ with the aim of slight performance deterioration. Examples include retrain with self-generated synthetic samples [9] or increased regularization, such as high weight decay [4].

4. **Attention-based methods** (e.g., SAG [119], PAG [7], SEG [118]) manually select and impair specific self-attention layers or attention maps. These strategies require trial and error, and they need to be tailored to each trained DM.

5. **Masked sliding window guidance** (M-SWG) aims at improving the global coherence of a sample. M-SWG [4] uses overlapping crops of the image or the latent representation that are processed independently from the positive models. Then, $\boldsymbol{\epsilon}_{\text{neg}}(\mathbf{x}_t)$ is defined as the aggregated outcome of the denoised crops.

Unlike CFG, the above methods apply to unconditional generation. Interestingly, strategies (1,4,5) do not require additional training or fine-tuning.

Alternative conditioning of generative models for label-free synthesis

Conditioning enables generative models to "specialize" in distinct groups of shared visual features [120, 80, 2]. Despite recent progress in DMs, the gap between ground-truth (GT) conditional and unconditional generation is significant. Unlike guidance methods [114], alternative conditions to the ground truth labels can be derived using auxiliary models, a process commonly referred to as *self-conditioning*. Self-conditioning signals can be

broadly categorized into image-level and sub-image-level types. Image-level conditioning applies a single condition to all pixels in an image, such as synthetic text captions [224]. Sub-image-level conditioning targets specific parts or regions of an image [120, 18]. For example, Hu et al. [120] extend DMs to incorporate conditioning based on bounding boxes or segmentation masks derived from auxiliary models.

When dealing with unlabelled data, learned representations from auxiliary models can play a pivotal role. Bordes et al. [35] condition DMs directly on image representations (*oracle features from real images*). However, this approach is limited to reconstructing training samples and cannot generate novel samples. Bao et al. [80] computed k-means cluster assignments using contrastively learned features from the training data, but did not match the performance of GT conditioning. Previously, instance-conditioned GANs combined ground-truth labels with features from each image's nearest-neighbor set [44].

The first label-free approach to achieve competitive performance to ground-truth conditioning leverages pre-trained self-supervised encoders [120]. Hu et al. [120] use k-means to assign pseudo-labels as conditions. This comes with the challenging choice of choosing the number of clusters (C). Adaloglou et al. [2] explore heuristics to derive a range of values for C. By carefully optimizing C and using pretrained encoders such as DINO [42], state-of-the-art FID scores in label-free synthesis were achieved on CIFAR10, CIFAR100 [151], and FFHQ [138]. At the same time, cluster conditioning demonstrates a training sample efficiency of a factor of 3 compared to unconditional generation.

Recently, Li et al. [165] develop a state-of-the-art two-stage framework for unconditional generation. In the first stage, a DM is trained to learn the feature distribution of the pretrained encoder. In this way, one can sample novel features, unlike [35]. In the second stage, another generative model is trained to generate images based on the features of the trained DM. Finally, another promising alternative is to leverage the zero-shot capability of CLIP to derive pseudo-labels. Ramesh et al. [218] condition DMs on language-guided image representations. Zhou et al. [320] propose a language-free framework that generates text features for images directly.

3. Contributions and References

Contributions

So far, existing approaches have been described. Building upon previous methods, the author has focused on developing novel techniques that adapt the learned representations to downstream tasks. In the following chapters in Part II, the following works are going to be presented:

1. **TEMI** [6] (Chapter 4) describes a clustering method with strong outperformance over applying k-means on image features. A small clustering head is trained using nearest neighbors with a novel objective that has a single bounded hyperparameter, which requires minimal to no tuning.

2. **PLP** [5] (Chapter 5) provides an alternative to CLIP models for visual OOD detection. Instead of simply computing top-1 image-text similarities, the in-distribution label names are used to predict pseudo-labels per sample. A linear layer is trained on these pseudo-labels, and all the logits are used to compute the OOD detection score. No tuning of the visual encoder, with promising scaling behavior as the model sizes increase up to 2B parameters, unlike existing training-free methods.

3. **C-EDM** [2] (Chapter 6) assesses whether cluster assignments from learned features can be used as conditioning in generative diffusion models, as an alternative to unconditional image generation. A strategy for estimating the optimal range of clusters for generative learning is demonstrated. Experimental results demonstrate a threefold training sampling efficiency compared to unconditional generation, with performance comparable to ground-truth class-conditioned models.

4. In Chapter 7, we conduct a comprehensive benchmark study of feature-based clustering methods at scale, disentangling the impact of the following data-related factors: i) class imbalance, ii) class granularity, iii) the presence of easily recognizable classes, and iv) the ability to capture multiple classes.

5. **ClusterMine** [3] (Chapter 8) is a novel OOD detection method that utilizes widely available text corpora for positive label mining using CLIP. ClusterMine is the first method to achieve state-of-the-art OOD detection performance without access to the ground truth in-distribution label names. It extracts positive concepts from a large text corpus by combining visual-only sample consistency (via clustering) and zero-shot image-text consistency. Experimental results reveal that ClusterMine achieves state-of-the-art performance in full-spectrum OOD detection.

6. **M-SWG** [4] (Chapter 9) is a novel, training-free diffusion sampling method. M-SWG upweights long-range spatial dependencies by guiding the diffusion model with itself by selectively restricting its receptive field. As a standalone sampling method, M-SWG achieves a superior Inception score (IS) compared to previous state-of-the-art training-free approaches, without introducing sample oversaturation. In conjunction with existing guidance methods, such as classifier-free guidance (CFG) [114], M-SWG reaches state-of-the-art Frechet distance.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Nikolaos Adaloglou, Tim Kaiser, Felix Michels, and Markus Kollmann. Rethinking cluster-conditioned diffusion models for label-free image synthesis. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 3603–3613, February 2025.

[3] Nikolaos Adaloglou, Diana Petrusheva, Mohamed Asker, Felix Michels, and Markus Kollmann. Clustermine: Robust label-free visual out-of-distribution detection via concept mining from text corpora. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, March 2026.

[4] Nikolas Adaloglou, Tim Kaiser, Damir Iagudin, and Markus Kollmann. Guiding a diffusion model with itself using sliding windows. In *36th British Machine Vision Conference, BMVC 2025, November 2025, Sheffield, UK*. BMVA, 2025.

[5] Nikolas Adaloglou, Felix Michels, Tim Kaiser, and Markus Kollmann. Adapting contrastive language-image pretrained (CLIP) models for out-of-distribution detection. *Transactions on Machine Learning Research*, 2024.

[6] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[7] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. *arXiv preprint arXiv:2403.17377*, 2024.

[8] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. 2017.

[9] Sina Alemohammad, Zhangyang Wang, and Richard G Baraniuk. Neon: Negative extrapolation from self-training improves image generation. *arXiv preprint arXiv:2510.03597*, 2025.

[10] Benedikt Alkin, Lukas Miklautz, Sepp Hochreiter, and Johannes Brandstetter. MIM-refiner: A contrastive learning boost from intermediate pre-trained masked image modeling representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[12] Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022.

[13] Anonymous. Diffuse and disperse: Image generation with representation regularization. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. under review.

[14] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 456–473. Springer, 2022.

[15] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8443–8452, 2021.

[16] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[17] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR, 2022.

[18] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[19] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

[20] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[21] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. 2022.

[22] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[23] Christopher J Bates and Robert A Jacobs. Efficient data compression in perception and perceptual memory. *Psychological review*, 127(5):891, 2020.

[24] Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):5418, 2020.

[25] Irad Ben-Gal. Outlier detection. In *Data mining and knowledge discovery handbook*, pages 117–130. Springer, 2010.

[26] Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. *Advances in Neural Information Processing Systems*, 36:58324–58345, 2023.

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[28] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[29] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[30] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[31] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023.

[32] Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models. *arXiv preprint arXiv:2510.18457*, 2025.

[33] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.

[34] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[35] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about, 2022.

[36] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[37] Jirayu Burapacheep and Yixuan Li. Your classifier can be secretly a likelihood-based OOD detector. *Transactions on Machine Learning Research*, 2024.

[38] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[39] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[40] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[41] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[42] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[43] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

[44] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

[45] Derrick Adrian Chan and Siphesihle Philezwini Sithungu. Evaluating the suitability of inception score and fréchet inception distance as metrics for quality and diversity in image generation. In *Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems*, pages 79–85, 2024.

[46] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[47] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[48] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. pages 16296–16305, 2021.

[49] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[50] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[51] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021.

[52] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. 2025.

[53] Tianzhe Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin David Haeffele, Rene Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. 2024.

[54] Tianzhe Chu Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin Haeffele, Rene Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. International Conference on Learning Representations (ICLR), 2024.

[55] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Metaclip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

[56] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[57] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[58] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13693–13702, 2021.

[59] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

[60] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR, 2023.

[61] Sylvain Delattre and Nicolas Fournier. On the kozachenko–leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185:69–93, 2017.

[62] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

[63] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[64] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*, 2018.

[65] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[66] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[67] Sander Dieleman. Diffusion is spectral autoregression, 2024.

[68] Sander Dieleman. Generative modelling in latent space, 2025.

[69] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

[70] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[71] Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1934–1943, 2023.

[72] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36:60878–60901, 2023.

[73] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022.

[74] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[75] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[76] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[77] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[78] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[79] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

[80] Bao et al. Why are conditional generative models better than unconditional ones? In *NeurIPS Workshop on Score-Based Methods*, 2022.

[81] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022.

[82] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2024.

[83] Abolfazl Farahani, Behrouz Pourshojae, Khaled Rasheed, and Hamid R Arabnia. A concise review of transfer learning. In *2020 international conference on computational science and computational intelligence (CSCI)*, pages 344–351. IEEE, 2020.

[84] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. 2020.

[85] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

[86] Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad Khorrami. Clipscope: Enhancing zero-shot ood detection with bayesian scoring. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5346–5355. IEEE, 2025.

[87] Artyom Gadetsky, Yulun Jiang, and Maria Brbic. Let go of your labels with unsupervised transfer. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 14382–14407. PMLR, 21–27 Jul 2024.

[88] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2023.

[89] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. 2020.

[90] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[91] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[92] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[93] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[94] Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. On partial prototype collapse in the dino family of self-supervised methods. *arXiv preprint arXiv:2410.14060*, 2024.

[95] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision, 2022.

[96] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[97] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[98] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8401–8409, 2019.

[99] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023.

[100] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[101] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[102] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[103] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 2022.

[104] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[105] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[106] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[107] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[108] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[109] Alvin Heng, Harold Soh, et al. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37:43952–43974, 2024.

[110] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[111] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[112] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[113] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[114] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[115] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[116] Laura Hollink, Aysenur Bilgin, and Jacco Van Ossenbruggen. Is it a fruit, an apple or a granny smith? predicting the basic level in a concept hierarchy. *arXiv preprint arXiv:1910.12619*, 2019.

[117] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey. *Int. J. Comput. Vision*, 133(10):7154–7200, July 2025.

[118] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 66743–66772. Curran Associates, Inc., 2024.

[119] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7462–7471, October 2023.

[120] Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. Self-guided diffusion models. In *CVPR*, 2023.

[121] Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[122] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[123] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. 2024.

[124] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

[125] Mark Ibrahim, David Klindt, and Randall Balestriero. Occam's razor for self supervised learning: What is sufficient to learn good representations? In *NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice*, 2024.

[126] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[127] Joebert S Jacaba. Audio compression using modified discrete cosine transform: The mp3 coding standard. *University of the Philippines, Manila*, 2001.

[128] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[129] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[130] Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *arXiv preprint arXiv:2506.08010*, 2025.

[131] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[132] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. 2022.

[133] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025.

[134] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025.

[135] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[136] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[137] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

[138] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[139] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International conference on machine learning*, pages 16049–16096. PMLR, 2023.

[140] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[141] Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *Scientific Reports*, 15(1):4876, 2025.

[142] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.

[143] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[144] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[145] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[146] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[147] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[148] Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. 2020.

[149] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025.

[150] Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. *arXiv preprint arXiv:2504.16064*, 2025.

[151] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[152] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[153] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[154] Gitaek Kwon, Jaeyoung Kim, Hong-Jun Choi, Byung-Moo Yoon, Sungchul Choi, and Kyu-Hwan Jung. Improving out-of-distribution detection performance using synthetic outlier exposure generated by visual foundation models. In *BMVC*, pages 10–11, 2023.

[155] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[156] Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025.

[157] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[158] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[159] Kyungeun Lee, Jaeill Kim, Suhyun Kang, and Wonjong Rhee. Towards a rigorous analysis of mutual information in contrastive learning. *Neural Networks*, 179:106584, 2024.

[160] Johannes Lehner, Benedikt Alkin, Andreas Fürst, Elisabeth Rumetshofer, Lukas Miklautz, and Sepp Hochreiter. Contrastive tuning: A little help to make masked autoencoders forget. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2965–2973, 2024.

[161] Jiachen Lei, Keli Liu, Julius Berner, Haiming Yu, Hongkai Zheng, Jiahong Wu, and Xiangxiang Chu. Advancing end-to-end pixel space generative modeling via self-supervised pre-training. *arXiv preprint arXiv:2510.12586*, 2025.

[162] Jinglun Li, Kaixun Jiang, Zhaoyu Chen, Bo Lin, Yao Tang, Weifeng Ge, and Wenqiang Zhang. Synthesizing near-boundary ood samples for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4496–4506, October 2025.

[163] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11578–11589, 2023.

[164] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. 2021.

[165] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2024.

[166] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024.

[167] Yicen Li, Haitz Sáez de Ocáriz Borde, Anastasis Kratsios, and Paul D. McNicholas. Keep it light! simplifying image clustering via text-free adapters. *CoRR*, abs/2502.04226, 2025.

[168] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *Forty-first International Conference on Machine Learning*, 2024.

[169] Jian Liang, Lijun Sheng, Zhengbo Wang, Ran He, and Tieniu Tan. Realistic unsupervised clip fine-tuning with universal entropy optimization. In *International Conference on Machine Learning (ICML)*, 2024.

[170] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[171] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[172] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[173] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[174] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[175] Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pages 22528–22538. PMLR, 2023.

[176] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[177] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Klaus Robert Muller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*, 2022.

[178] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[179] Alice V. De Lorenci, Seung Eun Yi, Théo Moutakanni, Piotr Bojanowski, Camille Couprie, Juan C. Caicedo, and Wolfgang Maximilian Anton Pernice. Scaling channel-adaptive self-supervised learning. *Transactions on Machine Learning Research*, 2025.

[180] Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58(2):1–39, 2025.

[181] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[182] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[183] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[184] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[185] Yu Mao, Yufei Cui, Tei-Wei Kuo, and Chun Jason Xue. Trace: A fast transformer-based general-purpose lossless compressor. In *Proceedings of the ACM Web Conference 2022*, pages 1829–1838, 2022.

[186] Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does CLIP's generalization performance mainly stem from high train-test similarity? In *The Twelfth International Conference on Learning Representations*, 2024.

[187] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

[188] Xiao-Li Meng and David Van Dyk. The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(3):511–567, 1997.

[189] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[190] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022.

[191] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022.

[192] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Helen Li, Ziwei Liu, Toshihiko Yamasaki, and Kiyoharu Aizawa. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *Transactions on Machine Learning Research*, 2025.

[193] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023.

[194] Sina Mohseni, Arash Vahdat, and Jay Yadawa. Shifting transformation learning for out-of-distribution detection. *arXiv preprint arXiv:2106.03899*, 2021.

[195] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

[196] Théo Moutakanni, Maxime Oquab, Marc Szafraniec, Maria Vakalopoulou, and Piotr Bojanowski. You don't need domain-specific data augmentations when scaling self-supervised learning. In *NeurIPS 2024-The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[197] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pages 311–320. Springer, 2018.

[198] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

[199] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[200] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[201] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[202] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[203] Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajic, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. *arXiv preprint arXiv:2509.16131*, 2025.

[204] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1695, 2023.

[205] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2023.

[206] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[207] Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[208] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEit v2: Masked image modeling with vector-quantized visual tokenizers. 2023.

[209] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022.

[210] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Representation Learning*, volume 2024, pages 1862–1874, 2024.

[211] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International conference on machine learning*, pages 5171–5180. PMLR, 2019.

[212] Qi Qian. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16645–16654, 2023.

[213] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[214] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[215] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[216] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[217] Nima Rafiee, Rahil Gholamipoor, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, and Markus Kollmann. Self-supervised anomaly detection by self-distillation and negative sampling. In *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings; Part IV*, pages 459–470. Springer, 2022.

[218] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[219] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[220] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[221] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[222] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

[223] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[224] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[225] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[226] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[227] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *ICLR*, 2024.

[228] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

[229] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[230] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

[231] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[232] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.

[233] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[234] Vikash Sehwag, Mung Chiang, and Prateek Mittal. {SSD}: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021.

[235] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. In *International Conference on Algorithmic Learning Theory*, pages 92–107. Springer, 2008.

[236] Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14410–14419, 2024.

[237] Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4k resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9631–9640, 2025.

[238] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025.

[239] Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*, 2025.

[240] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[241] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5484–5494, 2023.

[242] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

[243] Stefano Soatto and Alessandro Chiuso. Modeling visual representations: Defining properties and deep approximations. In *ICLR (Poster)*, 2016.

[244] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[245] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[246] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

[247] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[248] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[249] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[250] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.

[251] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[252] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

[253] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European conference on computer vision*, pages 691–708. Springer, 2022.

[254] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[255] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[256] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. 2023.

[257] Junjiao Tian, Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Exploring covariate and concept shift for out-of-distribution detection. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[258] Yonglong Tian, Olivier J Henaff, and Aäron Van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10063–10074, 2021.

[259] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[260] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[261] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[262] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015.

[263] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[264] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019.

[265] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[266] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023.

[267] Edy Umargono, Jatmiko Endro Suseno, and SK Vincensius Gunawan. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In *The 2nd international seminar on science and technology (ISSTEC 2019)*, pages 121–129. Atlantis Press, 2020.

[268] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[269] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[270] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7492–7501, 2022.

[271] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[272] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December 2010.

[273] Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, camille couprie, Maxime Oquab, Armand Joulin, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *Transactions on Machine Learning Research*, 2024.

[274] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[275] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[276] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[277] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

[278] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[279] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[280] Xi WANG, Nicolas Dufour, Nefeli Andreou, Marie-Paule CANI, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024.

[281] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023.

[282] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[283] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012, 2020.

[284] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? *CoRR*, abs/2202.08005, 2022.

[285] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[286] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[287] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8150–8159, 2019.

[288] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[289] Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying DINO via coding rate regularization. In *Forty-second International Conference on Machine Learning*, 2025.

[290] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023.

[291] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[292] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14475–14485, 2023.

[293] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024.

[294] Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025.

[295] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[296] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[297] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[298] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023.

[299] William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-OOD: Deciphering modern out-of-distribution detection algorithms. In *The Twelfth International Conference on Learning Representations*, 2024.

[300] Hui Yin, Amir Aryani, Stephen Petrie, Aishwarya Nambissan, Aland Astudillo, and Shengyuan Cao. A rapid review of clustering algorithms. *arXiv preprint arXiv:2401.07389*, 2024.

[301] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

[302] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.

[303] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.

[304] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning Diverse and Discriminative Representations via the Principle of Maximal Coding Rate Reduction, June 2020. arXiv:2006.08558 [cs].

[305] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[306] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022.

[307] Boxuan Zhang, Jianing Zhu, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han. What if the input is expanded in ood detection? *Advances in Neural Information Processing Systems*, 37:21289–21329, 2024.

[308] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017.

[309] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 102–117. Springer, 2020.

[310] Jiaqi Zhang, Juntuo Wang, Zhixin Sun, John Zou, and Randall Balestriero. Fastdinov2: Frequency based curriculum learning improves robustness and training speed. *arXiv preprint arXiv:2507.03779*, 2025.

[311] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2024.

[312] Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16650–16659, 2022.

[313] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[314] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[315] Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided OOD detection with vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[316] Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. *Advances in neural information processing systems*, 36:72110–72123, 2023.

[317] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. 2022.

[318] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[319] Xingzhi Zhou and Nevin L Zhang. Deep clustering with features from self-supervised pretraining. *arXiv preprint arXiv:2207.13364*, 2022.

[320] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022.

[321] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[322] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022.

Part II:
Core Publications

Exploring the Limits of Deep Image Clustering **94**

Adapting CLIP for Out-of-Distribution Detection **21**

Rethinking Cluster-Conditioned Diffusion Models **27**

Scaling Up Feature-Based Image Clustering Methods **26**

ClusterMine: Robust Label-Free Visual Out-Of-Distribution Detection via Concept Mining from Text Corpora **25**

Guiding a Diffusion Model with Itself using Sliding Windows **21**

Conclusion and Publication list **17**

Notes

- The asterisk symbol * denotes equal contribution. The author's name is highlighted in bold.
- Minimal modifications have been applied to the main manuscript. The author has removed parts of the appendices or integrated them into the main paper as needed.
- The papers are presented in chronological order as they were written to reflect the research trajectory of the author.
- The author has written additional publications during his PhD candidacy that are not included here. We provided a full list of publication at the end of the document. The aim is to emphasize the author's strongest scientific contributions, according to his own belief, and maintain a coherent flow between chapters.

4. Exploring the Limits of Deep Image Clustering using Pretrained Models

Outline

A Introduction 1
B Related Work 2
C Proposed Method 3
D Experimental Evaluation 7
E Conclusion 11
References 13
Appendix 17

Bibliographic Information

Adaloglou*, N., Michels*, F., Kalisch, H., & Kollmann, M. (2023). Exploring the limits of deep image clustering using pretrained models. Proceedings of the 34th British Machine Vision Conference (BMVC 2023), Aberdeen, UK, November 20–24, 2023. BMVA. `https://proceedings.bmvc2023.org/297/`

Corresponding author: *adaloglo@hhu.de*

Copyright Notice

©2023 The British Machine Vision Association and Society for Pattern Recognition. This is an accepted version of this article published in `https://doi.org/10.1007/978-3-031-15937-4_39`. Clarification of the copyright adjusted according to the guidelines of the publisher.

Acknowledgements

This work would not be possible without the help of Felix Michels (shared first-author), who contributed significantly in the prototyping phase as well as the mathematical proof

of Theorem 1.

Individual Contributions

The author proposed shifting the project's direction from unsupervised out-of-distribution detection on small-scale data to image clustering. The developed objective is based on pointwise mutual information, a concept primarily introduced and formulated by the author's supervisor. Within this context, the author contributed to prototyping the variation of pointwise mutual information designed explicitly for clustering. Following the emergence of competitive initial results, the author was responsible for developing the experiments and conducting a large part of them. Finally, the author was the leading writer of the first draft.

Abstract

We present a general methodology that learns to classify images without labels by leveraging pretrained feature extractors. Our approach involves self-distillation training of clustering heads based on the fact that the nearest neighbours in the pretrained feature space are likely to share the same label. We propose a novel objective that learns associations between image features by introducing a variant of pointwise mutual information together with instance weighting. We demonstrate that the proposed objective is able to attenuate the effect of false positive pairs while efficiently exploiting the structure in the pretrained feature space. As a result, we improve the clustering accuracy over k-means on 17 different pretrained models by 6.1% and 12.2% on ImageNet and CIFAR100, respectively. Finally, using self-supervised vision transformers, we achieve a clustering accuracy of 61.6% on ImageNet. The code is available at `https://github.com/HHU-MMBS/TEMI-official-BMVC2023`.

A Introduction

Given a plethora of publicly available pretrained vision models, we ask the following questions: a) how well-structured is the feature space of pretrained architectures with respect to label-related information, and b) how to best adapt this structure to unsupervised tasks. To answer these questions, we focus on unsupervised image classification, also known as image clustering. Image clustering is the task of assigning a semantic label to an image, given an a priori finite set of classes. Ultimately, image clustering consists of simultaneously learning the relevant representations and the cluster assignments.

To begin addressing the aforementioned questions, we present the key challenges regarding image clustering. First, given that we can roughly estimate the number of ground-truth labels, the underlying distribution among classes is hard to infer from the data, which is typically assumed to be uniform. Second, the images should ideally be classified consistently (images of the same class are grouped together) and confidently (one-hot prediction probability).

Consistency can be achieved by either learning features that are invariant under transformations of the same image (e.g. cropping, colour jitter, etc.) or invariant w.r.t. to substitution by other images that belong to the same semantic class. Consequently, clustering methods are generally prone to degenerate solutions [1]. In other words, samples tend to collapse into a single cluster, or the prediction probability spreads out uniformly.

It is well-established that representation learning plays a critical role in image clustering [6], which is achieved with self-supervised learning [7, 17, 41, 16]. Recent studies have demonstrated that self-supervised features are typically more transferable to new tasks than features from supervised learning [15]. The frequently used joint-embedding architectures [16, 5] are by design invariant to strong image transformations that preserve label-related information. That renders these architectures as promising candidates for image clustering [36], which has not been thoroughly explored at scale [42]. Even though self-supervision [15] and vision transformers (ViTs) [31] have been separately established for representation learning, limited research has been conducted to study self-supervised ViTs or vision-language models (i.e. CLIP [34]).

How to adapt a pretrained model for image clustering is non-trivial. For instance, it is well known that k-means is sub-optimal, as it often leads to imbalanced clusters [36] since it is primarily suitable for evenly scattered data samples around their centroids [39]. On the other hand, deep image clustering methods normally rely on pairs by mining the nearest neighbours (NN) based on their feature similarity [14, 20]. Still, images that are close in the feature space do not always share the same semantic class [36] and, therefore, must be considered as noisy pairs.

In this paper, a two-stage method that extends the existing multi-stage clustering approaches is proposed. In contrast to [36], where features are learned from scratch for each dataset, we show that multi-stage clustering approaches can leverage pretrained models trained on larger-scale datasets (Fig. 1) and focus on learning the cluster assignments. To this end, a self-distillation clustering framework is introduced using a novel objective based on pointwise mutual information and instance weighting. Second, a comprehensive experimental study across models and datasets is conducted. Therein, we report an average gain of 6.1% and 12.2% in clustering accuracy compared to k-means on ImageNet and CIFAR100 across 17 pretrained models, as illustrated in Fig. 4.1. Overall, we show

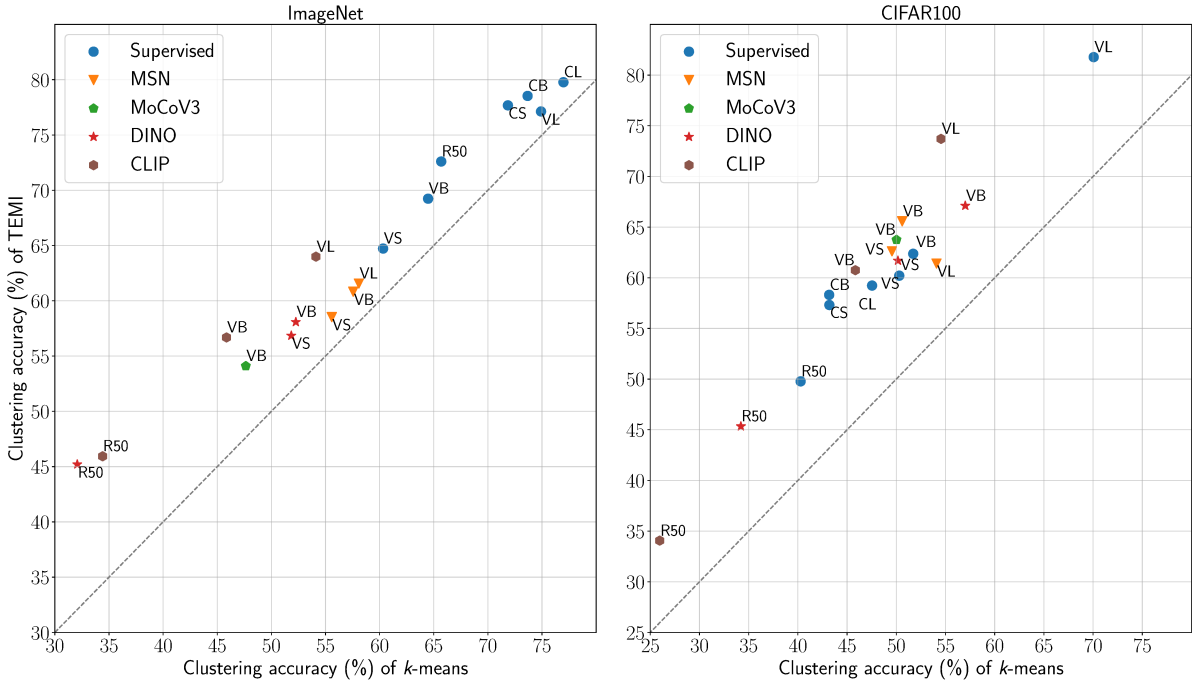


Figure 4.1: **Clustering accuracies on ImageNet (left) and CIFAR100 (right) across 17 pretrained models.** Supervised and self-supervised models (MSN, MoCoV3, DINO) were pretrained on ImageNet. R50 stands for ResNet50 [18], C for ConvNext [28], and V for Vision Transformer [13]. Small (S), Base (B), and Large (L) indicate the size of the models. The vertical distance of each data point to the diagonal (dashed line) shows the improvement of our method (TEMI) over k-means. Best viewed in color.

that ViTs capture the most transferable label-related features. Finally, we find that self-supervised ViTs [2] achieve state-of-the-art results (61.6% clustering accuracy) on ImageNet without using the ground-truth labels or external data.

B Related Work

B.1 Single-stage deep image clustering methods.

Deep image clustering approaches can be roughly divided into single and multi-stage methods. The majority of single-stage methods alternate between learning the features and the clusters, i.e. in an expectation-maximization manner. For instance, in DAC, [6] formulate a binary pairwise classification task, where pairs are selected based on their feature similarity at each iteration. Next, the computed pairs are used to train a convolutional neural network (CNN). In the same direction, in DeepCluster [4], the authors alternate between clustering the features of a CNN with k-means [29] and using the obtained cluster assignments as pseudo-labels to optimize the parameters of the CNN.

Later on, in [40], the authors demonstrate that DeepCluster is prone to degenerate solutions that are avoided via particular hyperparameter choices. To that end, the authors design a multi-step pseudo-label extraction framework called SeLa. The latter iteratively

estimates the pseudo-label assignment matrix under the equipartition constraint. In PCL [27], the authors formulate clustering as learning the cluster centroids with k-means in parallel with optimizing the network via contrastive learning [7]. To overcome the class collision of the negative pairs, [22] extend PCL in a proximal framework called ProPos. ProPos only maximizes the distance between the cluster centroids with contrastive learning while mining NN for neighbouring sample alignment [16]. However, most of the existing approaches still rely on k-means for estimating the clusters (pseudo-labels).

Several single-stage approaches exist, which aim to learn the feature representations and clusters jointly. Single-step methods are known to be sensitive to weight initialization [11]. In this direction, DCMM is developed [37] to progressively mine NN in the feature space as well as high-confident samples. Another single-stage end-to-end example is IIC, wherein [23] derives a mutual information-based objective for paired data to train a CNN. Nevertheless, the aforementioned approaches only consider stochastic transforms of the same image to obtain a pair. They are hence limited to solely learning invariances w.r.t. image augmentations, which cannot cover the variability of a given class [14]. More recently, [1] presented a single-stage end-to-end method, called SSCN, that employs a variant of the cross-entropy loss.

B.2 Multi-stage deep image clustering methods

Multi-stage methods initially design a pretext task in order to learn semantically meaningful features, such as denoising autoencoders [38]. A major breakthrough in deep image clustering is established by the adoption of contrastive learning [7, 17]. For instance, [36] decouples image clustering into three distinct steps, starting with contrastive learning. Subsequently, the authors train a head to cluster the mined NN from the extracted features. Lastly, they use the pseudo-labels from the confidently assigned samples to fine-tune the whole architecture. A similar approach, called NNM [11], aims to mine NN from the batch and dataset features.

Recently, TSP [42] leverage self-supervised pretrained ViTs [5] and train a clustering head, which is closer to our method. Nevertheless, their approach (TSP) heavily relies on k-means for the weight initialization phase. Surprisingly, very few image clustering approaches [36, 1, 40] have been successfully applied on ImageNet [12]. Besides, most methods report results only with Resnet50 [18], while superior architectures for image recognition remain unexplored [28, 13].

C Proposed Method

C.1 Classification model

Our aim is to learn a probabilistic classifier from pairs of examples that share label-related information. We assume that the data distribution, $p(x)$, is the result of a generative process, $c \sim p(c)$ and $x \sim p(x|c)$, with $p(c)$ the prior probability that an example belongs to a class $c \in \{1, .., C\}$. Consequently, the joint distribution, $p(x, x')$ that a pair of examples, (x, x'), belongs to the same class is given by $p(x, x') = \sum_{c=1}^{C} p(x|c)p(x'|c)p(c)$. We introduce a parametrized probabilistic classifier, $q(c|x)$, that distributes examples $x \sim p(x)$ among classes, with class occupancy given by $q(c) = \mathbb{E}_{x\sim p(x)}[q(c|x)]$. Using Bayes'

theorem $q(x|c) = q(c|x)p(x)/q(c)$, the joint distribution, $p(x, x')$, can be approximated by

$$q(x, x') = \sum_{c=1}^{C} q(x|c)q(x'|c)q(c). \tag{4.1}$$

To estimate the association between x and x' we introduce the pointwise mutual information, $\text{pmi}(x, x')$ [8], defined by

$$\text{pmi}(x, x') \coloneqq \log \frac{q(x, x')}{p(x)p(x')} = \log \sum_{c=1}^{C} \frac{q(c|x)q(c|x')}{q(c)}. \tag{4.2}$$

Theorem 1. *If (i) each example $x \sim p(x)$ belongs to one and only one cluster under the generative model $p(x) = \sum_c p(x|c)p(c)$, (ii) the joint distribution $p(x, x')$ is known, and (iii) $q^*(c|x)$ is a probabilistic classifier defined by*

$$q^*(c|x) = \arg\max_{q(c|x)} \mathbb{E}_{x,x' \sim p(x,x')}[\text{pmi}(x, x')], \tag{4.3}$$

then $q^*(c|x)$ is equal to the optimal probabilistic classifier, $p(c|x) = p(x|c)p(c)/p(x)$, up to a permutation of cluster indices.

The proof can be found in the appendix. Theorem 1 states that under condition (i) the knowledge of pairs of examples belonging to the same class suffices to establish an objective for an optimal classification model.

C.2 Self-distillation clustering framework

The starting point is a pretrained feature extractor (backbone) $g(\cdot)$ that assigns each example x in the dataset D a feature vector $g(x)$. We mine the k nearest neighbours (k-NN) of x in the feature space by computing the cosine similarity between $g(x)$ and the feature vectors of all other images in D. We denote the set of k-NN for x by S_x. During training, we randomly sample x from D along with x' from S_x, to generate image pairs that share label information with high probability.

We introduce two clustering heads, a *student head*, $h_s(\cdot)$, and a *teacher head*, $h_t(\cdot)$, that share the same architecture but differ w.r.t. their parameters, θ_s, and θ_t. Each head consists of a three-layer fully connected feed-forward network. The image pairs x, x' are fed to the shared backbone and subsequently, in the two heads, $h_s(g(x))$ and $h_t(g(x'))$. The head outputs are converted to probabilistic classifiers, $q_s(c|x)$ and $q_t(c|x')$, using a temperature-scaled softmax function, which for the student's head is given by

$$q_s(c|x) = \frac{\exp(h_s(g(x))_c \,/\, \tau)}{\sum_{c'} \exp(h_s(g(x))_{c'} \,/\, \tau)}, \tag{4.4}$$

where τ is the temperature hyperparameter. Unlike previous self-distillation frameworks [5], we use the same temperature $\tau = 0.1$ for both heads. We approximate the pointwise mutual information by

$$\widetilde{\text{pmi}}(x, x') \coloneqq \log \sum_{c=1}^{C} \frac{q_s(c|x)q_t(c|x')}{\tilde{q}_t(c)} \tag{4.5}$$

and estimate $q(c)$ by an exponential moving average (EMA) over batches using the teacher's head

$$\tilde{q}_t(c) \leftarrow m\,\tilde{q}_t(c) + (1-m)\frac{1}{B}\sum_{i=1}^{B} q_t(c|x_i), \tag{4.6}$$

with B the batch size and $m \in (0,1)$ a momentum parameter. In practice, we symmetrize Eq. (4.5) to compute the loss function

$$\mathcal{L}(x,x') := -\frac{1}{2}\left(\widetilde{\mathrm{pmi}}(x,x') + \widetilde{\mathrm{pmi}}(x',x)\right). \tag{4.7}$$

Note that only the parameters θ_s of the student's head are updated using backpropagation. The parameters of the teacher's head, θ_t, are updated by an exponential moving average for the student parameters, θ_s, over past update steps [5, 16]. As a result, $p_t(c|x)$ represents a sufficiently stable target distribution for the student head. In contrast to other self-distillation frameworks [5], no complicated adaptation of softmax temperatures over training is required. Following previous work [36], we employ an ensemble of H independent clustering heads in training (Fig. 4.2), which alleviates the variability stemming from random initialization. We use the teacher head with the lowest training loss for the evaluation.

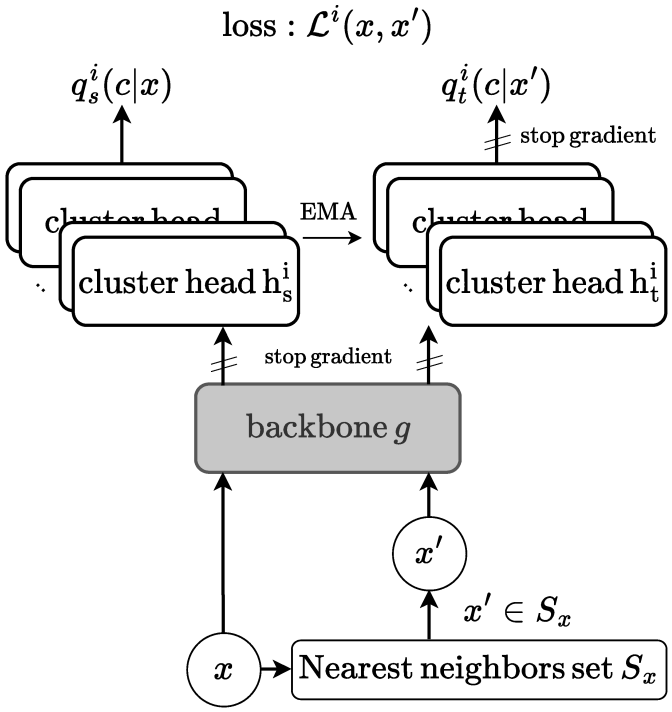


Figure 4.2: **An overview of the proposed self-distillation clustering framework.** The nearest neighbors are mined in the feature space of g. EMA refers to the exponential moving average over the parameters.

C.3 Balancing class utilization

For a dataset D that has been generated using balanced classes, $p(c) = const$, we expect that $\tilde{q}(c) \approx const$, as a consequence of the optimization process. However, in practice, classes are typically far from uniformly utilized. We suspect that our self-distillation learning framework leads to over-confident class predictions for a fraction of classes in the early training phase.

To limit the effect of aligning the cluster predictions and take into account the individual cluster probability of the mini-batches $\tilde{q}_t^i(c)$, we introduce a hyperparameter $\beta \in (0.5, 1]$ in Eq. (4.5) to avoid collapsing all sample pairs in a single cluster. Without affecting the optimal solution, we rewrite Eq. (4.5) as

$$\widetilde{\text{pmi}}^{\,i}(x, x') = \log \sum_{c=1}^{C} \frac{(q_s^i(c|x) q_t^i(c|x'))^\beta}{\tilde{q}_t^i(c)}, \tag{4.8}$$

where $i \in \{1, \ldots, H\}$ is the head index.

Intuition for β.

We now explain how the above equation addresses the discussed challenges of image clustering. Eq. (4.8) consists of two parts inside the log sum: the numerator $(q_s^i(c|x) q_t^i(c|x'))^\beta$ encourages the class assignment of a positive pair to align (consistency) and is maximal when this assignment is one-hot. The denominator $\tilde{q}_t^i(c)$ promotes a uniform cluster distribution by up-weighing the summand corresponding to classes with low probability ($\tilde{q}_t^i(c)$ is low). In effect, β balances these two effects by reducing the influence of the numerator, thus avoiding degenerative solutions.

Note that for $\beta = 0.5$, the loss corresponds to the Bhattacharyya distance [3] if $\tilde{q}_t^i(c) = const$. The Bhattacharyya distance can be minimal even if q_t^i is far from one-hot. Moreover, if utilization of all classes is not required, for example, as in the case of overclustering, we set $\beta = 1$. Crucially, β is bounded, and we empirically found that the value of 0.6 consistently avoids degenerative solutions across the majority of datasets, as opposed to existing clustering methods [36, 1]. In addition, we propose an experimental strategy to choose β without access to the ground-truth labels, as explained in Section D.2. The symmetrized loss from Eq. (4.8) is defined as $\mathcal{L}^i(x, x')$ in analogy to Eq. (4.7).

C.4 Teacher-guided instance weighting

As discussed in Section A, the mined k-NN in the feature space of $g(\cdot)$ tend to be noisy. For this reason, we introduce an instance weighting term for each head i given by

$$w_i(x, x') = \sum_{c=1}^{C} q_t^i(c|x) q_t^i(c|x'). \tag{4.9}$$

Intuitively, $w_i(x, x')$ acts as a guidance term that assigns a higher weight to true positive pairs compared to false positive ones. Importantly, $w_i(x, x')$ relies only on the predictions of the teacher. The rationale behind this is that model averaging over training iterations tends to produce more accurate predictions [35, 33]. We call this setup teacher-weighted

pointwise mutual information (WPMI). The final objective for each separate head i is given by $\mathcal{L}^i_{\text{WPMI}}(x, x') \coloneqq w_i(x, x')\mathcal{L}^i(x, x')$.

TEMI: Teacher ensemble-weighted pointwise mutual information

To compensate for the noisy NN pairs based on the feature space of g, we further propose aggregating information from multiple heads, in contrast to previous works [36] that employ independent heads. For this purpose, we compute a scalar for each image pair using the mean weight across the heads, which is conceptually similar to model ensembling. We thus call this loss TEMI (teacher ensemble-weighted pointwise mutual information) defined by

$$\mathcal{L}^i_{\text{TEMI}}(x, x') \coloneqq \frac{1}{H}\sum_{j=1}^{H} w_j(x, x')\mathcal{L}^i(x, x'). \tag{4.10}$$

D Experimental Evaluation

D.1 Datasets, metrics and implementation details

The proposed method (TEMI) is evaluated on five common benchmark datasets, namely CIFAR10, CIFAR20, CIFAR100 [25], STL10 [9], and ImageNet [12]. CIFAR10, CIFAR20 and CIFAR100 contain $50K$ training images, STL10 contains $5K$ training samples, and ImageNet has 1281167 training samples. CIFAR20 has the same training data as CIFAR100 with 20 superclasses derived from the ground-truth labels. We resize all images to 224×224. The training set is used during the optimization phase, while the evaluations are carried out on the validation set.

We report the clustering accuracy (ACC) and the adjusted random index (ARI) to quantify the clustering performance. To estimate the accuracy, the one-to-one mapping between cluster predictions and ground-truth labels is computed by the Hungarian matching algorithm [26]. For our overclustering experiments, we report the adjusted mutual information (AMI). Finally, we establish two baselines: a) k-means and b) the SCAN clustering loss within our self-distillation framework. For a fair comparison with existing methods, we assume to know the number of ground-truth labels in advance.

Concerning the hyperparameters, we set $H = 50$ and $\beta = 0.6$ for clustering, while we set $\beta = 1$ for overclustering. We use 25-NN on ImageNet and 50-NN for the remaining datasets. We used the AdamW optimizer [30] for 200 epochs with a batch size of 512, with a learning rate of 10^{-4}, weight decay of 10^{-4}, and report results at the end of training. Unlike previous methods [36], we found that augmentations (RandAugment [10], and the ones from [7]) were not improving the clustering metrics when training with k-NN pairs. Hence, we precomputed the feature representations, which enables training the clustering heads on a single GPU with 12GB of VRAM within 24 hours.

D.2 Experimental Results

We first present a strategy to choose $\beta \in (0.5, 1]$. As depicted in Fig. 4.3, an accurate model, $p_t(c|x)$ should be able to maintain a high entropy $\text{H}(q_t(c))$, while maintaining its discriminative power. To quantify the latter, we use the conditional entropy $\text{H}(q_t(c|x))$.

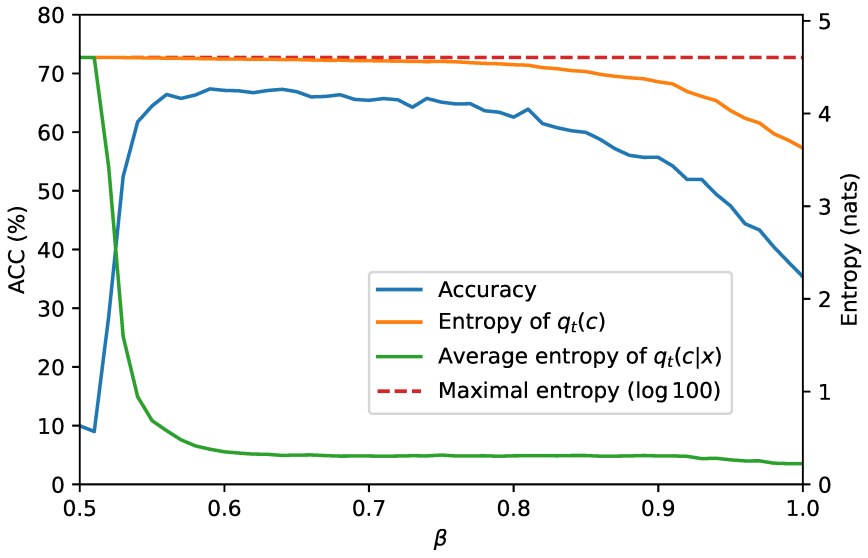


Figure 4.3: **Effect of β on the validation accuracy and on the entropy of $q_t(c|x)$ and $q_t(c)$ on CIFAR100**. The values are computed using TEMI DINO ViT-B/16. The dashed horizontal line illustrates the maximal possible entropy, i.e. $\log C$. A high entropy of $q_t(c)$ indicates that the clusters are almost uniformly utilized, while a low entropy of $q_t(c|x)$ indicates highly confident predictions (one-hot).

The lower the value of $\mathrm{H}(q_t(c|x))$, the more discriminative the predictions. The extreme case $\mathrm{H}(q_t(c|x)) = 0$ corresponds to a one-hot distribution. Thus, we propose to pick the lowest value of β such that $\mathrm{H}(q_t(c|x))$ remains sufficiently low. We experimentally found 0.6 to work consistently well across models and datasets.

As shown in Table 4.1, an average accuracy gain of 5.0% over k-means is found for CIFAR100 and ImageNet, even with the plain PMI setup. Introducing multiple heads in PMI further improves the obtained results by an average gain of 0.8%. Critically, for our best setup (TEMI), we observe an average gain of 8.1% and 3.7% compared to k-means and the SCAN clustering loss, respectively. Note that even 16 heads were sufficient to get similar performance, specifically less than 1% accuracy deterioration compared to 50 heads.

To study the applicability of our method, we then applied our best setup (TEMI) to various publicly available pretrained models, as shown in Fig. 4.1. Therein, we report an average accuracy gain of 6.1% and 12.1% compared to k-means on ImageNet and CIFAR100 across 17 different pretrained models. More specifically, TEMI MSN ViT-L/14 and TEMI DINO ViT-B/16 are the best-performing self-supervised methods on ImageNet (61.6% ACC) and CIFAR100 (67.1% ACC). Moreover, CLIP-based backbones have the highest ACC increase over k-means when trained with TEMI, precisely 10.7% on ImageNet and 14.1% on CIFAR100.

Concerning the supervised pretrained models in Fig. 4.1, we demonstrate that ConvNext-L outperforms ViT-L on ImageNet, precisely by 2.7% on ACC with TEMI. However, the supervised ViT-L surpasses ConvNext-L by a large margin of 22.3% in ACC when benchmarked on CIFAR100 with TEMI. Large ViTs learn the most transferable label-related features among the architectures investigated, even without supervised

Method	Heads	CIFAR100	ImageNet
k-means	-	56.99	52.26
SCAN	50	62.6±0.94	55.6±0.15
PMI	1	61.6±0.41	57.5±0.22
WMI	1	63.4±1.89	56.5±0.41
PMI	50	63.1±0.56	57.7±0.06
WPMI	50	65.6±1.04	57.0±0.38
TEMI	50	**67.1±1.30**	**58.4±0.22**

Table 4.1: **Ablation study for the TEMI objective**. All the experiments were conducted with $\beta = 0.6$, and DINO ViT-B/16 as the backbone model. The clustering accuracy is reported in %.

Method	Architecture	NMI (%)	ACC (%)	ARI (%)
SeLa [40]	Resnet50	65.7	30.5	16.2
SCAN [36]	Resnet50	72.0	39.9	27.5
SSCN [1]	Resnet50	73.3	41.1	29.5
Our method				
TEMI (DINO pretraining)	Resnet50	74.5	45.2	31.3
TEMI (DINO pretraining)	ViT-B/16	81.4	58	45.9
TEMI (MSN pretraining)	ViT-L/16	**82.5**	**61.6**	**48.4**

Table 4.2: **Unsupervised image classification (clustering) results for the ImageNet validation set, without using the ground-truth labels or additional data**. Evaluation metrics include clustering accuracy (ACC), normalized mutual information (NMI), and adjusted random index (ARI). All our models are pretrained on ImageNet.

fine-tuning. This finding is consistent with [31].

Clustering results on ImageNet. Regarding ImageNet, we compare various self-supervised architectures that were trained without any external data, as depicted in Table 4.2. Using the same architecture (Resnet50) as current state-of-the-art models (SSCN [1]), TEMI achieves an improvement of 4.1% in ACC. With MSN ViT-L/16 as the backbone, we push the state-of-the-art ACC on ImageNet to 61.6%, resulting in a substantial gain of 20.5% compared to SSCN. The obtained results strongly indicate that first learning the augmentation-invariant features and then focusing on learning the invariances w.r.t. images that belong to the same class is an effective strategy for image clustering.

Overclustering results. Incentivized by the above observation, we investigate the overclustering performance in Table 4.3 by adopting the setup from [27]. More concretely, we use 25K clusters and set $\beta = 1$ as in ProPos. We almost match the performance of ProPos [22] with TEMI DINO Resnet50 without tuning the number of clusters or any other hyperparameter while reaching a considerable gain of 7.4% in AMI with TEMI DINO ViT-B/16.

Small-scale benchmarks. In Table 4.4, the transfer performance on three small-

Method	AMI (%)
DeepCluster [4]	28.1
MoCo [17]	28.5
PCL [27]	41.0
ProPos [22]	52.5
TEMI DINO Resnet50	51.8±0.1
TEMI DINO ViT-B/16	**59.9±0.2**
TEMI MSN ViT-L/16	58.8±0.5

Table 4.3: **Overclustering results on the ImageNet validation set.** The adjusted mutual information (AMI) score for 25K clusters is reported, as in [27]. For all experiments, we set $\beta = 1$.

scale datasets is evaluated. TEMI DINO ViT-B backbone has inarguably the best transfer performance, outperforming the ACC of ProPos by 4.6% and TSP by 2.9% on average. It is worth pointing out that TSP [42] uses the same pretrained model, thus a fair comparison. Ultimately, we notice a large accuracy gap between clustering methods and probing in CIFAR20, which suggests that the superclass structure cannot be inferred from the visual input. For instance, clocks, lamps, and telephones are grouped into household electrical devices.

Analysis on noise (false positives) from the NN of g. As shown in Tab. 5, when keeping only the true positive neighbors from 50-NN, we increased the performance from 67.1 $\rightarrow$ 82.6 using TEMI DINO ViT-B on CIFAR100. We also show that the head weighting term of TEMI in Eq. 10 is not needed, highlighting that TEMI is designed for the noisy pairs obtained from k-NN. As a reference, DINO ViT-Base has 72% true positive pairs in 20-NN and 66% in the mined 50-NN on CIFAR100.

D.3 Discussion

How expressive can an image classifier be by only training with NN pairs? We examine the training accuracy in Table 4.5, by training with the true positive pairs from the computed k-NN. The 98.6% training accuracy on CIFAR100 with TEMI DINO ViT-B/16 indicates that it is possible to train a powerful unsupervised image classifier by relying only on pairs. In fact, we observe that we almost match the supervised linear probing accuracy on CIFAR100 (84.1% vs 85.3%). Still, we identify cases where the human-annotated label is ambiguous and cannot be determined solely by the visual signal.

What is the impact of the instance weighting term? After training, we examined the actual value of the instance weighting term $w(x, x')$. To this end, we computed the mean weights for true positives and false positives sampled from 50-NN within the CIFAR100 validation set, which take the values 0.76 and 0.40, respectively. Furthermore, $w(x, x')$ has a negative impact when only true positive pairs are considered during training (Table 4.5). This is an expected behavior, as a fraction of true positive pairs will be down-weighted by $w(x, x')$ due to low feature similarity (i.e. digital and analog clocks).

How discriminative are the cluster assignments of TEMI? Besides Fig. 2, we

	CIFAR10			CIFAR20			STL10		
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
DAC [6]	39.6	52.2	30.6	18.5	23.8	8.8	36.6	47	25.7
DCCM [37]	49.6	62.3	40.8	28.5	32.7	17.3	37.6	48.2	26.2
PICA [21]	59.1	69.6	51.2	31	33.7	17.1	61.1	71.3	53.1
NNM [11]	74.8	84.3	70.9	48.4	47.7	31.6	69.4	80.8	65
PCL [27]	80.2	87.4	76.6	52.8	52.6	36.3	71.8	41.0	67.0
SCAN [36]	79.7	88.3	77.2	48.6	50.7	33.3	69.8	80.9	64.6
SPICE [32]	86.5	92.6	85.2	56.7	53.8	38.7	87.2	93.8	87.0
ProPos* [22]	88.6	94.3	88.4	60.6	61.4	45.1	75.8	86.7	73.7
TSP† [42]	88.0	94.0	87.5	61.4	55.6	43.3	95.8	97.9	95.6
TEMI DINO ViT-B/16†	**88.6**	**94.5**	**88.5**	**65.4**	**63.2**	**48.9**	**96.5**	**98.5**	**96.8**
TEMI MSN ViT-L/16†	82.9	90.0	80.7	59.8	57.8	42.5	93.6	96.7	93.0
(Natural language) supervised pretraining									
TEMI CLIP ViT-L/14†	92.6	96.9	93.2	64.5	61.8	46.8	96.4	97.4	94.9
TEMI Sup. ViT-L/16†	91.8	96.0	91.6	65.0	58.4	45.4	82.7	84.6	73.9
Supervised baselines									
Probing DINO ViT-B/16†	92.5	96.8	93.1	82.4	89.5	79.5	97.8	99.2	98.2
Probing MSN ViT-L/16†	91.5	96.4	92.3	80.7	88.2	77.0	96.8	98.8	97.4
Probing CLIP ViT-L/14†	95.1	98.1	95.8	85.7	91.7	83.6	99.2	99.7	99.4
Probing Sup. ViT-L/16†	91.5	96.5	92.4	83.7	90.8	81.7	98.0	99.3	98.4

Table 4.4: **Clustering performance metrics (in %) on small-scale benchmark datasets, evaluated on their validation splits.** Probing means training a linear layer on top of the pretrained backbone in a supervised manner. We only highlight the best self-supervised pretrained model as the new state-of-the-art. We clarify that methods with † use models pretrained on external data, while * indicates methods that include additional dataset splits during training (i.e. validation data).

quantify the discriminative power of TEMI by computing the mean and median maximum softmax probability (MSP [19]). We calculate a mean and median MSP of 88.5% and 98.9% on CIFAR100 and 85.3% and 99.2% on ImageNet (Fig. 4.4). The computed results verify that the introduced framework results in discriminative predictions.

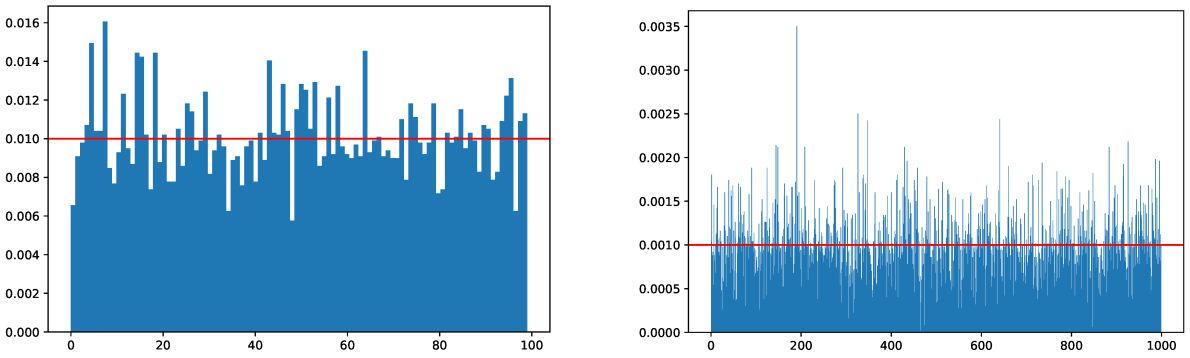
Joint learning of encoder and cluster head. When jointly training the pretrained backbone with the already trained head, we observed a performance increase only when the pretraining dataset was different from the downstream dataset (67.1% $\rightarrow$ 70.9% ACC on CIFAR100 using the ImageNet-pretrained DINO ViT-B/16). We hypothesize this enables learning features specific to the training distribution. Still, the structure in the latent space of the pretrained model is required for TEMI to determine the k-NN.

E Conclusion

In this paper, a novel and general self-distillation framework for image clustering that can achieve competitive results almost out of the box was proposed. In addition, a new objective based on pointwise mutual information was presented. After studying the performance of 17 pretrained models, it was shown that TEMI can be used with any pretraining with significant improvements over k-means. Finally, new state-of-the-art

Loss	Val. ACC	Train ACC
PMI	**84.1 ± 0.36**	**98.6 ± 0.38**
TEMI	82.6 ± 0.67	96.5 ± 0.88
supervised baseline		
Probing	85.3	99.3

Table 4.5: **Clustering accuracies on CIFAR100 when training only with the true positive NN pairs using TEMI DINO ViT-B/16.**



(a) CIFAR100 using TEMI DINO ViT-B/16 (b) ImageNet using TEMI MSN ViT-L/16

Figure 4.4: **Histogram of cluster assignments on different datasets**. The horizontal red line illustrates the ideal histogram, where all clusters would be uniformly utilized. We also compute the KL divergence between the predictions and the uniform distribution on CIFAR100 and ImageNet, which is $1.5 \cdot 10^{-2}$ and $5 \cdot 10^{-2}$, respectively. The predictions would be uniform in the extreme case where the KL divergence is 0.

results were achieved on ImageNet both for clustering and overclustering, leveraging self-supervised ViTs. To conclude, future works are encouraged to explore the connection between image clustering and representation learning in greater depth.

References

[1] Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022.

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 456–473. Springer, 2022.

[3] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[6] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[11] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13693–13702, 2021.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[15] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[20] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *International Conference on Machine Learning*, pages 2849–2858. PMLR, 2019.

[21] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8849–8858, 2020.

[22] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[23] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[26] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[27] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[29] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[30] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[31] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[32] Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022.

[33] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[35] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[36] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[37] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8150–8159, 2019.

[38] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[39] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pages 3861–3870. PMLR, 2017.

[40] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

[41] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[42] Xingzhi Zhou and Nevin L Zhang. Deep clustering with features from self-supervised pretraining. *arXiv preprint arXiv:2207.13364*, 2022.

Datasets	ImageNet 50			ImageNet 100			ImageNet 200		
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
SCAN	82.2	76.8	66.1	80.8	68.9	57.6	77.2	58.1	47.0
Propos	82.8	-	69.1	83.5	-	63.5	80.6	-	53.8
TEMI DINO	86.10	80.01	70.93	85.65	75.05	65.45	85.20	73.12	62.13
TEMI MSN	**88.14**	**84.87**	**76.46**	**88.53**	**82.86**	**74.08**	**86.65**	**77.96**	**66.70**
(natural language) supervised pretraining									
TEMI CLIP	92.32	88.27	82.78	90.06	83.43	75.81	88.39	77.76	69.41
TEMI Sup.	95.75	95.12	91.40	94.95	92.50	87.95	93.94	90.37	84.05

Table 4.6: **Clustering performances on ImageNet subsets.** All subsets were evaluated on their respective validation splits, as detailed in Table 5.2.

Appendix

App. I Further discussion points

Additional computational complexity from multiple heads. In theory, the computational time complexity of TEMI by adding multiple heads is linear, given a sequential implementation. In practice, due to GPU-related optimizations, we find it to be faster. In fact, training on a single Nvidia A100 GPU takes only 4 GB of memory with 50 heads on CIFAR100. Training takes just about 45 minutes because we precompute the feature representations, while training with just one head takes about 5 minutes.

config	value
optimizer	AdamW
base learning rate	10^{-4}
weight decay	10^{-4}
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	512, 1024 (ImageNet)
learning rate schedule	constant
softmax temperature τ	0.1
β	0.6, 0.55 (CIFAR20)
cluster heads	50
warmup epochs	20, 10 ImageNet
training epochs	200, 800 (STL10)
teacher momentum	0.996
augmentation	None

Table 4.7: **Hyperparameters for training the clustering heads.**

A Note on CIFAR100 VS CIFAR20. We observe that previous works have established the CIFAR20 as a clustering benchmark. However, we believe that the CIFAR20 superclasses are not an ideal benchmark for image clustering. In the reported

results in Table 4.4, one can easily notice that all models perform worse in CIFAR20 than in CIFAR100. Examples that justify the performance gap include a) clocks, computer keyboards, lamps, telephones, and televisions are grouped into household electrical devices, b) bridges, castles, houses, roads, and skyscrapers are grouped into large man-made outdoor things, and c) bears, leopards, lions, and wolfs are grouped into carnivores. These examples illustrate that the superclasses are not separable from the pixel information alone.

Additional implementation details and results on ImageNet-based subsets. To enforce reproducibility, the mean performance metrics are reported for all our experiments and metrics, computed over 3 independent runs with different seeds. For a fair comparison with SCAN, we tune its entropy regularization hyperparameter, λ, based on a grid search and use the value $\lambda = 4$. Crucially, we found that some pretrained models (i.e. MSN) produce unnormalized features. For that reason, we standardize the features of all models before feeding them to the clustering heads. For the linear probing experiments, we trained a linear layer using the Adam [24] optimizer with a learning rate of 10^{-3} and weight decay of 10^{-3}. Finally, Table 4.6 presents results on the ImageNet subsets proposed by SCAN.

App. II Proof of Theorem 1

We will first show that

$$q^*(c|x) = \arg\max_{q(c|x)} \mathbb{E}_{x,x' \sim p(x,x')}[\mathrm{pmi}(x, x')], \tag{4.11}$$

is bounded and leads to the correct joint distribution.

Lemma 1. *The mutual information*

$$\mathrm{I}(x; x') = \int p(x, x') \log \frac{p(x, x')}{p(x)p(x')} \mathrm{d}x\mathrm{d}x'$$

is an upper bound for the expected pointwise mutual information. In particular,

$$\begin{aligned} &\mathbb{E}_{x,x' \sim p(x,x')}[\mathrm{pmi}(x, x')] \\ =&\mathrm{I}(x; x') - \mathrm{KL}(p(x, x') \parallel q(x, x')), \end{aligned}$$

where KL *is the Kullback–Leibler divergence.*

Proof.

$$\begin{aligned} & \mathbb{E}_{x,x' \sim p(x,x')}[\mathrm{pmi}(x, x')] && (4.12)\\ = & \int p(x, x') \log \frac{q(x, x')}{p(x)p(x')} \mathrm{d}x\mathrm{d}x' && (4.13)\\ = & \int p(x, x') \log \left(\frac{p(x, x')}{p(x)p(x')} \frac{q(x, x')}{p(x, x')} \right) \mathrm{d}x\mathrm{d}x' && (4.14)\\ = & \int p(x, x') \log \frac{p(x, x')}{p(x)p(x')} \mathrm{d}x\mathrm{d}x' && (4.15)\\ & - \int p(x, x') \log \frac{p(x, x')}{q(x, x')} \mathrm{d}x\mathrm{d}x' && (4.16)\\ = & \mathrm{I}(x; x') - \mathrm{KL}(p(x, x') \parallel q(x, x')). && (4.17) \end{aligned}$$

□

For the proof of Theorem 1, we now assume that each example x belongs to exactly one cluster c and need to show that the model $q^*(c|x)$ maximizing the objective

$$q^*(c|x) = \arg\max_{q(c|x)} \mathbb{E}_{x,x' \sim p(x,x')}[\mathrm{pmi}(x, x')]$$

is equal to $p(c|x)$ up to a permutation of the clusters.

Proof. Since $p(c|x)$ is one-hot by assumption, let us denote the class to which an image x belongs as c_x, so that we have

$$p(c|x) = [c = c_x],$$

where the Iverson bracket $[c = c_x]$ is 1 if $c = c_x$ and 0 otherwise. We denote the prediction of the classifier by $\hat{c}_x$, with

$$\hat{c}_x \coloneqq \operatorname*{argmax}_c q^*(c|x).$$

An equivalent formulation of the theorem then is: $q^*(c|x)$ is one-hot for every x and $\hat{c}_x = \hat{c}_{x'}$ if and only if $c_x = c_{x'}$.

Using the same factorization we used to formulate the pointwise mutual information in Equation (2) we obtain

$$\frac{p(x, x')}{p(x)p(x')} = \sum_{c=1}^{C} \frac{p(c|x)p(c|x')}{p(c)} = [c_x = c_{x'}]p(c_x)^{-1}.$$

Lemma 1 already states that the objective is maximized if and only if $q^*(x, x') = p(x, x')$ and therefore

$$\begin{aligned}\mathrm{pmi}(x, x') &= \sum_{c=1}^{C} \frac{q^*(c|x)q^*(c|x')}{q^*(c)} \\ &= \frac{q^*(x, x')}{p(x)p(x')} = \frac{p(x, x')}{p(x)p(x')} = [c_x = c_{x'}]p(c_x)^{-1}.\end{aligned}$$

If $c_x \neq c_{x'}$, we have

$$0 \leq q^*(\hat{c}_x|x)q^*(\hat{c}_x|x')/q^*(\hat{c}_x) \leq \mathrm{pmi}(x, x') = 0.$$

Since $q^*(\hat{c}_x|x) > 0$, this implies $q^*(\hat{c}_x|x') = 0$ and therefore $\hat{c}_x \neq \hat{c}_{x'}$. Furthermore, from the pigeonhole principle it follows that $q^*(c|x) = 0$ for $c \neq c_x$ which both implies that $q^*(c|x)$ is one-hot as well as $\hat{c_x} = \hat{c}_{x'}$ if $c_x = c_{x'}$, therefore concluding the proof. □

App. III Randomly sampled images from TEMI cluster assignments

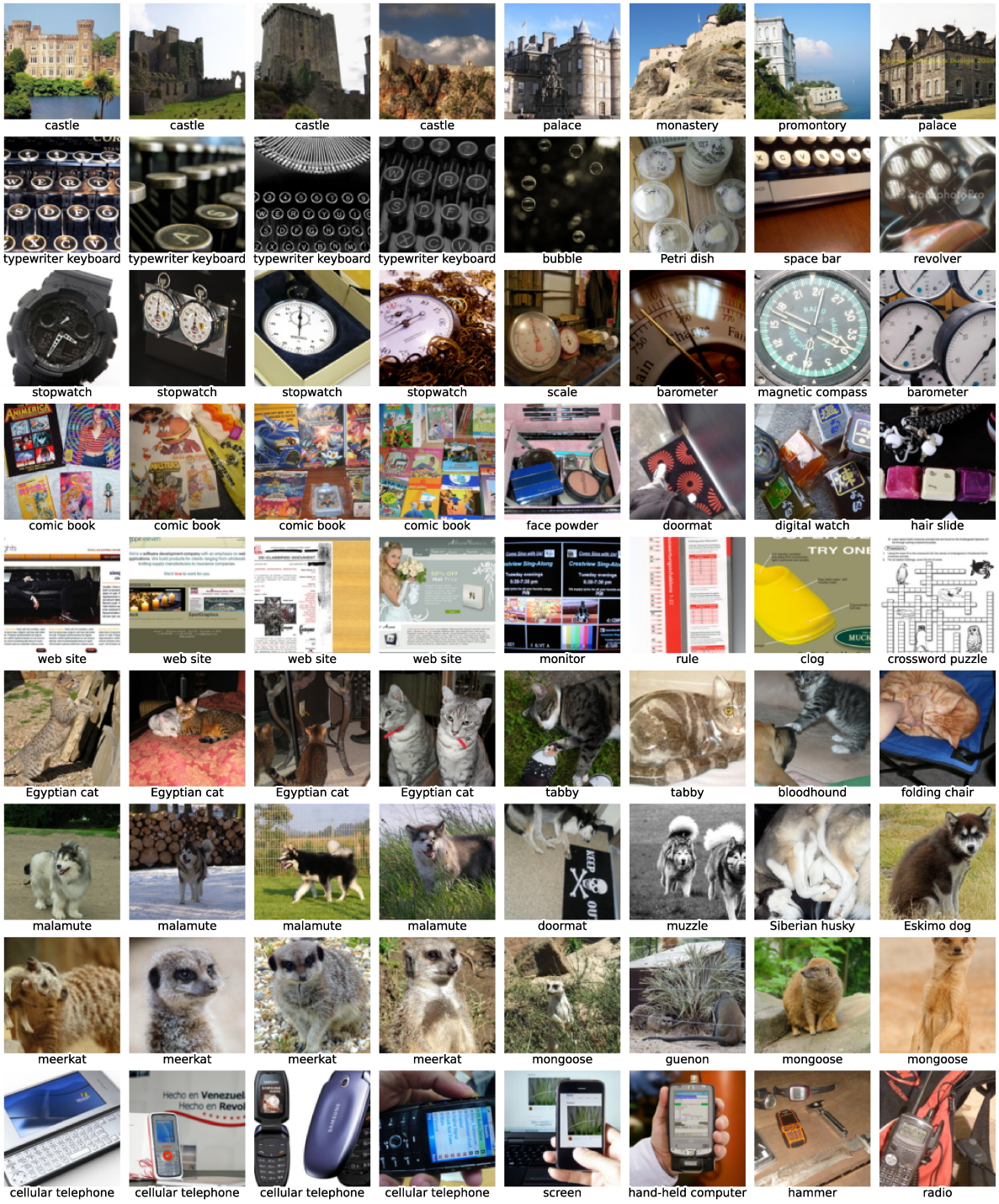


Figure 4.5: **Randomly sampled images from the ImageNet dataset that are assigned in the same cluster using the TEMI MSN ViT-L/16 model**. The ground-truth label is indicated in the text under the image. The images in each row are assigned to the same cluster. The first four columns correspond to correctly classified images, while the last four are examples of misclassified images.

5. Adapting Contrastive Language-Image Pretrained (CLIP) Models for Out-of-Distribution Detection

Outline

A Introduction 1
B Related Work 3
C Proposed OOD Detection Setup 4
D Experimental Results 10
E Discussion 13
F Conclusion 14
References 16
Appendix 24

Bibliographic Information

Adaloglou, N., Michels, F., Kaiser, T., & Kollmann, M. (2024). Adapting Contrastive Language-Image Pretrained (CLIP) Models for Out-of-Distribution Detection. Transactions on Machine Learning Research.

Corresponding author: *adaloglo@hhu.de*

Copyright Notice

©2024 BY-SA 4.0 (Attribution-ShareAlike 4.0), Transactions on Machine Learning Research. This is an accepted version of this article published in `https://openreview.net/forum?id=YCgX7sJRF1`. Clarification of the copyright was adjusted according to the guidelines of the publisher.

Acknowledgements

We would like to thank Georgios Zoumpourlis and Kaspar Senft for their insightful remarks and discussions on the manuscript. The author would additionally like to thank Felix Michels (second co-author) for i) the development of the feature-based adversarial OOD data manipulation method presented in this work, and ii) his invaluable help during the rebuttal period on reproducibility concerns raised by the reviewers regarding the baseline methods.

Individual Contributions

Following the emergence of vision-language pretrained models (commonly referred to as CLIP), the author, together with his supervisor, hypothesized that CLIP could serve as an effective out-of-distribution (OOD) detector. Under this prism, the author discovered that, even without fine-tuning, CLIP models exhibit a strong positive correlation between in-distribution classification accuracy and unsupervised OOD detection performance (i.e. AUROC). Building on this observation, concurrently with developing exercise materials for the MSc course on representation learning, Felix Michels and the author identified a simple yet effective variant that integrates zero-shot inference with linear (pseudo-)probing. This approach yielded promising preliminary results. Subsequently, the author extended and scaled the analysis of these preliminary results covering: i) supervised, ii) unsupervised, and iii) label name-based OOD detection. Finally, the author served as the lead writer of the manuscript.

Abstract

We present a comprehensive experimental study on pre-trained feature extractors for visual out-of-distribution (OOD) detection, focusing on leveraging contrastive language-image pre-trained (CLIP) models. Without fine-tuning on the training data, we are able to establish a positive correlation ($R^2 \geq 0.92$) between in-distribution classification and unsupervised OOD detection for CLIP models in 4 benchmarks. We further propose a new simple and scalable method called *pseudo-label probing* (PLP) that adapts vision-language models for OOD detection. Given a set of label names of the training set, PLP trains a linear layer using the pseudo-labels derived from the text encoder of CLIP. Intriguingly, we show that without modifying the weights of CLIP or training additional image/text encoders (i) PLP outperforms the previous state-of-the-art on all 5 large-scale benchmarks based on ImageNet, specifically by an average AUROC gain of 3.4% using the largest CLIP model (ViT-G), (ii) linear probing outperforms fine-tuning by large margins for CLIP architectures (i.e. CLIP ViT-H achieves a mean gain of 7.3% AUROC on average on all ImageNet-based benchmarks), and (iii) billion-parameter CLIP models still fail at detecting feature-based adversarially manipulated OOD images. The code is available at `https://github.com/HHU-MMBS/plp-official-tmlr2024`.

A Introduction

Transferring the representations of pretrained vision models has improved the performance on a plethora of image recognition tasks [86, 74, 61, 1]. To date, these models are trained with various types of supervision, which accelerates training convergence compared to random initialization [28]. Examples include self-supervision [10], natural language supervision [63], weakly-supervised learning [53], or standard supervised learning. Concurrently, vision transformers (ViTs) have been established [17], along with an enormous number of variants [48, 75, 5], as a suitable architecture for training large-scale models in the visual domain [15].

Nevertheless, the applicability of the learned features of pretrained models is crucial and non-trivial, especially for unsupervised downstream tasks [7, 1]. This work centers on adapting pretrained feature extractors for various visual OOD detection setups, focusing on contrastive language-image pretraining (*CLIP*) models.

The task of OOD, novelty, or anomaly detection aims at identifying whether a given test sample is drawn from the *in-distribution* (the training set) or an alternative distribution, known as the *out-distribution*. Accurate detection of anomalies is indispensable for real-world applications to ensure safety during deployment [2, 67]. The detected unfamiliar samples can be processed separately, possibly with a human expert in the loop, rather than making a potentially uncalibrated prediction [27]. Despite significant advances in deep learning, neural networks tend to generate systematic errors for test examples far from the training set [59] or assign higher likelihoods to OOD samples compared to in-distribution samples [57].

Recent studies have established a firm connection between the training data distribution accuracy and OOD generalization [31, 78, 15]. A similar connection has been identified for supervised OOD detection when in-distribution training labels are available for training or fine-tuning. Supervised training leads to intermediate representations that likely form tight label-related clusters [21]. An ideal representation for OOD detection should capture semantic properties, such as the pose and shape of an object, while remaining sensitive to the properties of the imaging process (e.g., lighting, resolution) [80].

A suitable choice of visual feature representations is critical for detecting anomalies. However, learning informative representations for unsupervised OOD detection [73, 70, 64], where no in-distribution labels are available, is a challenging and active research area [14]. Unsupervised methods often adopt self-supervision to learn the in-distribution features by defining pretext tasks such as rotation prediction [24]. A major milestone in visual representation learning was reached by contrastive [10] and non-contrastive visual self-supervised methods [8]. Recently, CLIP has enabled learning from vast amounts of raw text [63]. Vision-language models offer the unprecedented advantage of zero-shot image classification by leveraging the label names of the downstream dataset. Nonetheless, adapting CLIP models is not straightforward, even in the supervised scenario. For instance, Pham et al. [62] showed that fine-tuning CLIP models degrades their robustness against classifying distribution-shifted samples.

Labeled OOD samples are typically unavailable in real-world applications, and the number of in-distribution samples is usually limited. Therefore, external data have been widely employed [64] in two ways: a) outlier exposure where the external data is treated as

anomalous [35], and b) using models pretrained on auxiliary data [71]. Outlier exposure leads to performance gains only if the auxiliary data are sufficiently diverse and disjoint from the in-distribution [35, 50]. On the other hand, [34, 31] showed that pretrained backbones can enhance OOD detection performance and robustness without relying on dataset-specific shortcuts [23]. Consequently, pretrained models are suitable candidates for OOD detection, while [22] has recently established CLIP models for OOD detection, especially when in-distribution label names are present [55].

In parallel, several OOD detection methods still rely on similar small-scale benchmarks based on low-resolution images [56, 64, 20], such as CIFAR [44]. Huang et al. [39] argued that methods explicitly tuned for these benchmarks may not always translate effectively into larger-scale and real-life applications. Towards this direction, new large-scale and more challenging benchmarks have been introduced [37, 84, 6], which consider ImageNet [16] as in-distribution. Finally, even though the robustness against adversarial attacks has been sufficiently explored in image classification [72, 25, 54], less attention has been given to studying the construction of robust visual OOD detectors [4, 85]. Even though several advancements in visual feature extractors have been made and new large-scale OOD detection benchmarks [6] have been proposed, limited research has been conducted in OOD detection regarding the choice of pretrained model and evaluation scheme, especially regarding CLIP models [55, 22, 20].

In this paper, we present an experimental study across 25 feature extractors and several visual OOD detection benchmarks. Using the existing publicly available models, we demonstrate that large-scale CLIP models are robust unsupervised OOD detectors and focus on adapting the representations of CLIP under different OOD detection settings. Under this scope, we examine several OOD detection setups based on the availability of labels or image captions (i.e. in-distribution class names). The core contributions of this work are summarized as follows:

- To investigate whether the dependence between in-distribution accuracy and unsupervised OOD detection performance can be confirmed without in-distribution fine-tuning, we quantify the performance of 25 pretrained models across 4 benchmarks (Fig. 5.2). CLIP models exhibit the strongest positive correlation without fine-tuning across all benchmarks (R^2 coefficient ≥ 0.92). Interestingly, the features of CLIP ViT-G outperform the ones from supervised ImageNet pretraining on ImageNet-based OOD detection.

- To adapt the representations of CLIP for OOD detection, we propose a simple and scalable method called pseudo-label probing (PLP). The text-based pseudo-labels are computed using its text encoder based on the maximum image-text feature similarity. We leverage the obtained text-based pseudo-labels to train a linear layer on top of CLIP. Without modifying the weights of CLIP or training separate text/image encoders, PLP surpasses the previous state-of-the-art [55] on 5 ImageNet benchmarks by an average AUROC gain of 1.8% and 3.4% for CLIP ViT-H and CLIP ViT-G, respectively. Moreover, linear probing achieves superior performance on ImageNet-based OOD benchmarks compared to fine-tuning, notably by a mean AUROC gain of 7.3% using CLIP ViT-H.

- Finally, we show that a simple feature-based adversarial attack tailored for OOD

detection can easily fool CLIP ViT-G trained on billions of samples (86.2% $\rightarrow$ 50.3% AUROC deterioration) by changes that are invisible to humans.

B Related Work

B.1 Supervised OOD detection methods

Supervised OOD detection methods rely on the fact that in-distribution classification accuracy is positively correlated with OOD detection performance [21, 22]. For that reason, many OOD detection methods derive anomaly scores from supervised in-distribution classifiers. Hendrycks et al. [33] developed the maximum softmax probability (MSP) as OOD detection score, which is frequently used, or its temperature scaled version [46]. More recently developed scores based on the logits of the in-distribution classifier are the maximum logit [30] or the negative energy scores by [47]. An alternative OOD detection score, which requires the in-distribution labels, is the parametric Mahalanobis-based score [45]. The Mahalanobis score assumes that the representations from each class are normally distributed around the per-class mean and conform to a mixture of Gaussians [66, 21]. Later on, [71] established an important yet simple OOD detection score, namely the k-nearest neighbors (NN) distance, without requiring the feature norms [73] or temperature tuning [64]. The k-NN distance has the advantage of being non-parametric and model- and distribution-agnostic.

Supervised learning may not always produce sufficiently informative features for identifying OOD samples [80]. To this end, additional tasks have been proposed to enrich the supervised-learned features. Examples include determining the key in-distribution transformations and predicting them [36] or contrastive learning. Mohseni et al. [56] attempt to first learn the domain-specific transformations for each in-distribution using Bayesian optimization. Zhang et al. [87] present a two-branch framework, where a generative flow-based model and a supervised classifier are jointly trained.

B.2 Unsupervised OOD detection methods

Unsupervised OOD detection methods rely on learning in-distribution features, typically accomplished with contrastive or supervised contrastive learning [70, 42]. Contrastive-based methods can be further enhanced by designing hand-crafted transformations that provide an estimate of near OOD data [64] or by developing better OOD detection scores [73]. For example, Tack et al. [73] add a transformation prediction objective (i.e. rotation prediction such as $[0°, 90°, 180°, 270°]$). In SSD [70], the authors define a simpler contrastive-based OOD detection method, where the Mahalanobis distance is computed in the feature space using the cluster centers of k-means [51].

B.3 OOD detection methods using external data or pretrained models

Early works attempted to use external data to generate examples near the OOD decision boundary or incorporate them for outlier exposure and negative sampling [35, 64]. Nonetheless, disjointness between the in and out-distribution cannot be guaranteed,

and applying shifting transformations depends on the in-distribution [56], which limits the applicability of such approaches.

Hendrycks et al. [34] show that large-scale models pretrained on diverse external datasets can boost OOD detection performance. Recent OOD detection scores, which leverage pretrained models, deal with the large semantic space by, for instance, grouping images with similar concepts into small groups as in [39]. The majority of existing methods focus on supervised OOD detection [71, 22]. Such approaches include fine-tuning the whole network or parts of it [65]. Contrarily, Ren et al. [66] showed that a Mahalanobis-based score could achieve comparable OOD detection performance on small-scale benchmarks without fine-tuning.

Out of the limited label-free methods based on pre-trained models, a simple approach is the k-NN feature similarity, which has not been studied systematically across pre-trained models [71]. Another unsupervised approach aims at initially detecting an *a priori* set of visual clusters. The obtained clusters are subsequently used as pseudo-labels for fine-tuning [14]. Apart from unsupervised OOD detection, CLIP models can additionally leverage in-distribution class names [20], and in-distribution prototypes can be obtained using the textual encoder, which has not been thoroughly investigated. Esmaeilpour et al. [20] extend the CLIP framework by training a text-based generator on top, while Ming et al. [55] compute the maximum softmax probability of image-text feature similarities. To date, CLIPN [77] is the current large-scale OOD state-of-the-art using CLIP by training a separate text encoder on 3 million auxiliary text-image pairs. Still, it is unknown whether this approach scales on larger-scale CLIP models. Approaches that work with off-the-shelf-features without requiring external image-text data [77], fine-tuning, or training of separate encoders [77] remain relatively unexplored, especially in conjunction with CLIP models [81].

B.4 OOD detection robustness

Hendrycks et al. [32] analyzed the robustness under corruptions and geometric perturbations, such as Gaussian noise and brightness shift. Since it is not always clear which manually perturbed images are present in the in-distribution (and which are not), attention has been given to adversarial robustness [9]. Existing works have primarily focused on fooling supervised OOD detection methods on small-scale benchmarks [85]. Even though the zero-shot classification performance of CLIP deteriorates significantly when the input images are constructed adversarially [54], their adversarial OOD detection robustness has not been thoroughly investigated. Finally, even though adversarial feature-based methods exist [26, 38, 52], they have not been applied to OOD detection.

C Proposed OOD Detection Setup

C.1 Considered pretrained models

Several supervised CNNs, such as ResNet50 [29], ConvNext [49] and ViT [17] models trained on ImageNet and ImageNet-21K [16, 68] were utilized. Regarding Imagenet-pretrained self-supervised models, the DINO [8], MoCov3 [11], and MSN [3] were selected. Finally, CLIP-based models were either trained on OpenAI-400M [63] or LAION-2B [69],

	Images/second	Number of Params (M)
ResNet50	2719	24
ConvNext-S	1576	49
ConvNext-B	1157	88
ConvNext-L	695	196
ConvNext-B wide	888	88
ConvNext-L deep	531	200
ConvNext-XXL	180	847
ViT-S/16	3210	22
ViT-B/16	1382	86
ViT-L/16	475	303
ViT-H/14	183	631
ViT-G/14	78	1843

Table 5.1: **Number of parameters (in millions) and inference time (images per second on a single GPU) for the utilized network architectures.**

which consists of 400M and 2 billion image-text description pairs, respectively. Further information regarding the network complexities is reported in Table 5.1. To quantify the complexity, we performed inference on a single GPU with a batch size of 256 to compute the images per second that were processed at 224×224 resolution.

C.2 Datasets and metrics

We denote the in-distribution as $\mathcal{D}_{\text{in}}$, and the out-distribution as $\mathcal{D}_{\text{out}}$. The corresponding train and test splits are indicated with a superscript. To design a fair comparison between vision and vision-language models, we define *unsupervised OOD detection* without having access to the set of $\mathcal{D}_{\text{in}}$ label names. Following [39], we use ImageNet as $\mathcal{D}_{\text{in}}$ for large-scale OOD detection benchmarks and CIFAR for small-scale benchmarks.

For the large-scale benchmarks, we use the following 5 OOD datasets: Imagenet-O (IN-O) [37], a flower-based subset of iNaturalist [76], Texture [13], a subset of the SUN scene database [82], a subset of Places [88]. IN-O contains 2K samples from ImageNet-21K, excluding ImageNet. It is worth noting that Places, Textures, and IN-O have a class overlap of 59.5%, 25.6%, and 20.5% with ImageNet, according to [6]. Therefore, we additionally use the newly proposed NINCO [6], which contains 5879 samples belonging to 64 classes with zero class overlap with ImageNet's classes. Dataset information is summarized in Table 5.2.

To quantify the OOD detection performance, the area under the receiver operating characteristic curve (AUROC) and the false positive rate at 95% recall (FRP95) are computed between $\mathcal{D}_{\text{out}}^{test}$ test and $\mathcal{D}_{\text{in}}^{test}$. Below, we present the used OOD detection scores, given a pretrained backbone model g.

1-NN. For the unsupervised evaluations, we use the maximum of the cosine similarity [71] between a test image x' and $x_i \in \mathcal{D}_{\text{in}}^{\text{train}} = \{x_1, x_2 \ldots, x_N\}$ as an OOD score:

$$s_{\text{NN}}(x') = \max_i \text{sim}(g(x'), g(x_i)), \tag{5.1}$$

Dataset	Classes	Train images	Validation images
Pretraining datasets			
ImageNet	1K	1.28M	-
ImageNet-21K	21K	14M	-
OpenAI-400M	-	400M	-
LAION-2B	-	2B	-
In-distribution datasets			
CIFAR10	10	50K	10K
CIFAR100	100	50K	10K
ImageNet	1K	1.28M	50K
Out-distribution datasets			
iNaturalist	110	-	10K
SUN	50	-	10K
Places	50	-	10K
IN-O	200	-	2K
NINCO	64	-	5.88K
Texture	47	-	5.54K
CIFAR10-A	10	-	1000
CIFAR10-AS	10	-	1000

Table 5.2: **An overview of the number of classes and the number of samples on the considered datasets.**

where $\mathrm{sim}(\cdot)$ is the cosine similarity and N the number of $\mathcal{D}_{\text{in}}^{\text{train}}$ samples .

Mahalanobis distance (MD). The MD can be either applied directly on the feature space of the pre-trained model, $z_i = g(x_i)$ or on the trained linear head, $z_i = h(g(x_i))$. However, MD assumes that the in-distribution labels $y_i \in \{y_1, \ldots, y_N\}$ are available. We denote the class index $c \in \{1, \ldots, C\}$, with C being the number of $\mathcal{D}_{\text{in}}$ classes and N_c the number of samples in class c. For each class c, we fit a Gaussian distribution to the representations z [45]. Specifically, we first compute the per-class mean $\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} z_i$ and a shared covariance matrix

$$\Sigma = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} (z_i - \mu_c)(z_i - \mu_c)^\top. \tag{5.2}$$

The Mahalanobis score is then computed for each test sample as

$$\mathrm{MD}_c(z') = \big(z' - \mu_c\big)\Sigma^{-1}\big(z' - \mu_c\big)^\top, \tag{5.3}$$

$$s_{\mathrm{MD}}(x') = -\min_c \mathrm{MD}_c(z')\,. \tag{5.4}$$

MD can also be applied with cluster-wise means, for instance, using the k-means cluster centers computed on the feature space of g [70]. We denote this score as k-means MD and use the number of ground truth $\mathcal{D}_{\text{in}}$ classes to compute the cluster centers.

Relative Mahalanobis distance (RMD). Given the in-distribution mean $\mu_0 = \frac{1}{N}\sum_i^N z_i$, we additionally compute $\Sigma_0 = \frac{1}{N}\sum_i^N (z_i - \mu_0)(z_i - \mu_0)^\top$ to compute MD_0

analogously to Eq. (5.3). Subsequently, the RMD score [66] can be defined as

$$s_{\text{RMD}}(x') = -\min_c \big(\text{MD}_c(z') - \text{MD}_0(z')\big)\,. \tag{5.5}$$

Energy. As in [47], negative free energy is computed over the logits z of an $\mathcal{D}_{\text{in}}$ classifier as

$$s_{\text{energy}}(x') = T \cdot \log \sum_i^C e^{z'_i/T}, \tag{5.6}$$

where $T = 1$ is a temperature hyperparameter.

C.3 Pseudo-Label Probing (PLP) using CLIP's textual encoder

The pseudo-MSP baseline. Ming et al. [55] proposed a simple visual OOD detection method to leverage the text encoder of CLIP by feeding the $\mathcal{D}_{\text{in}}$ class names in parallel with computing the image representations using the image encoder of CLIP. A set of captions such as "a photo of a $\{label\}$" is used to produce the text representations. The text-based representations are averaged after $L2$ normalization. Subsequently, $L2$ normalized text-image cosine similarities for each test image are computed. The maximum softmax probability (MSP) is used as an OOD score for text-image similarities. This is conceptually similar to using MSP on the logits of a supervised $\mathcal{D}_{\text{in}}$ classifier. We use pseudo-MSP as a baseline.

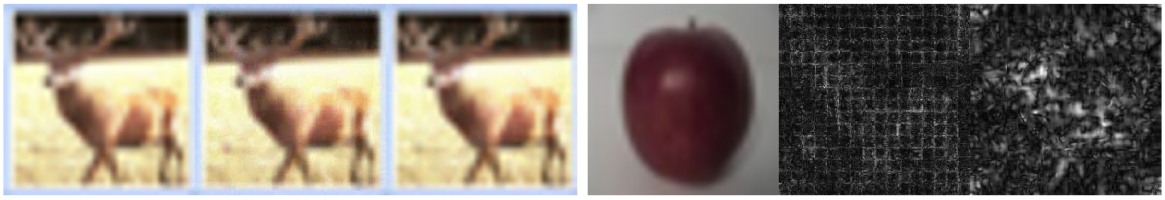
PLP. We propose a simple method called *pseudo-label probing* (PLP) that extends pseudo-MSP. After computing the image-text similarity, we derive a pseudo-label for each $\mathcal{D}_{\text{in}}^{train}$ image using $argmax$, the class index that corresponds to the maximum image-text similarity. To enhance the predictions of CLIP to assign high probabilities to $\mathcal{D}_{\text{in}}^{train}$, a linear layer is then trained with the obtained pseudo-labels on the features of g. During the OOD detection evaluation stage, the RMD or energy score is used, namely *PLP + RMD* and *PLP + Energy*.

Similar to pseudo-MSP, PLP remains OOD data-agnostic and assumes no prior information on $\mathcal{D}_{\text{out}}$. PLP is conceptually similar to self-training [83]. Differently from standard self-training, we only train a linear layer using the text-based prototype that is closest to each $\mathcal{D}_{\text{in}}$ image in feature space derived from the textual encoder of CLIP. Finally, while both pseudo-MSP and PLP do not modify the parameters of CLIP or train external image/text encoders [20, 77], we show that PLP scales better with larger-scale models (Section D) while training a linear layer adds minimal overhead compared to pseudo-MSP.

We additionally consider supervised linear probing as an alternative to the existing paradigm of fine-tuning visual feature extractors. Linear probing refers to training a linear head on the features of the backbone g, using the $\mathcal{D}_{\text{in}}^{\text{train}}$ labels and acts as an upper bound for PLP. Subsequently, RMD or Energy is computed. This is in contrast to existing approaches [21, 39] that typically fine-tune the visual backbone, which is significantly slower, computationally more expensive, and may lead to catastrophic forgetting [41].

C.4 Feature-based adversarial OOD data manipulation

The unsupervised OOD detection performance of CLIP ViT-G, combined with its known classification robustness against natural distribution shifts, raises the question of whether



(a) (b) (c) (d) (e) (f)

Figure 5.1: **Generating an adversarial OOD example of a deer that is close enough in the feature space to an in-distribution image of an apple**. From left to right: a) the original OOD image from the CIFAR10 test set, b) the adversarial example without smoothing, c) the adversarial example with smoothing, d) the randomly sampled in-distribution target image from CIFAR100, e) the per-pixel Euclidean distance between the original image and perturbed image, and f) the distance between the original and the smoothly perturbed image.

these models are also adversarially robust OOD detectors. To expose the vulnerability of CLIP as an OOD detector, we develop a feature-based adversarial data manipulation tailored for OOD detection similar to [26, 38, 52]. We solely aim to test adversarial OOD detection; thus, we do not compare it with feature-based methods that can be modified for OOD detection.

For a test image $x' \in \mathcal{D}_{\text{out}}^{\text{test}}$, we randomly choose an in-distribution image $x \in \mathcal{D}_{\text{in}}^{\text{train}}$ as the target. Instead of maximizing the OOD scores directly [4, 9], we create an adversarial perturbation ρ with the same dimensions as x' that maximizes the cosine similarity between the in-distribution feature $g(x)$ and $g(x' + \rho)$. We use the Adam optimizer to compute ρ by minimizing $-\operatorname{sim}(g(x), g(x' + \rho))$, starting with Gaussian noise $\rho \sim \mathcal{N}(0, 10^{-3})$ and clipping $x' + \rho$ to the pixel range $[0, 1]$ after every update step, similar to [85]. We emphasize that we do not directly restrict the perturbation size and only limit the number of steps, as opposed to [85].

We experimentally observe that in the case of ViTs, the perturbations are quite visible along the edges of the transformer patches (Fig. 5.1). To create more natural appearing adversarial examples, we enforce the smoothness of the perturbation by regularizing the allowed perturbation difference between neighboring pixels. We compute the image gradient $\partial\rho/\partial h$ and $\partial\rho/\partial w$ in the horizontal and vertical direction, respectively. The image gradients have the same shape as the image, $3 \times H \times W$, and we define the regularization term as

$$\ell_{\text{smooth}}(\rho) = \frac{1}{3HW} \sum_{ijk} \left(\frac{\partial \rho}{\partial h} \right)_{ijk}^2 + \left(\frac{\partial \rho}{\partial w} \right)_{ijk}^2, \tag{5.7}$$

where i, j, k run over image dimensions. We then minimize the loss

$$\ell_{\text{adv}} = -\operatorname{sim}(g(x), g(x' + \rho)) + \lambda \ell_{\text{smooth}}(\rho), \tag{5.8}$$

with respect to the pertubation ρ, where λ is a hyperparameter. During the evaluation, we remove the chosen target image, x, from $\mathcal{D}_{\text{in}}^{\text{train}}$ to show that the adversarial example, $x' + \rho$, cannot be detected as OOD from the remaining in-distribution examples. As a proof of concept, we create two adversarial OOD datasets[1] for the CIFAR100 $\rightarrow$ CIFAR10

[1] `https://drive.google.com/drive/folders/1pYGEPQwagRzdKlPqMQv7sRH6C9MV5vNF`

Method	OOD Dataset									
	iNaturalist Plants		SUN		IN-O		Texture		NINCO	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIP ViT-B										
Ming et al. [55]	**30.91**	**94.61**	**37.59**	**92.57**	-	-	57.77	86.11	-	-
Pseudo-MSP†	52.10	89.55	40.34	91.92	78.20	79.47	55.28	**86.82**	83.09	72.44
PLP + Energy	47.70	93.36	58.15	88.69	**79.00**	**80.72**	**54.63**	86.44	**81.93**	**75.65**
CLIP ViT-L										
Ming et al. [55]	28.38	94.95	**29.00**	**94.14**	-	-	59.88	84.88	-	-
Pseudo-MSP†	48.37	90.98	29.85	93.91	72.15	81.49	57.92	85.42	74.72	78.47
PLP + Energy	**27.24**	**95.40**	44.99	91.05	**64.70**	**85.81**	**46.75**	**88.22**	**71.95**	**83.83**
CLIP ViT-H (LAION-2B)										
Pseudo-MSP†	60.35	90.45	42.92	91.50	59.65	87.00	46.95	89.67	74.24	82.99
PLP + Energy	**48.61**	**93.02**	**40.90**	**91.78**	**52.50**	**89.24**	**35.00**	**91.96**	**71.9**	**84.88**
CLIP ViT-G (LAION-2B)										
Pseudo-MSP†	56.49	90.55	38.69	92.75	55.50	88.11	46.95	89.70	71.30	83.38
PLP + Energy	**23.51**	**96.01**	**34.41**	**92.90**	**42.80**	**91.39**	**30.38**	**92.74**	**59.32**	**88.65**

Table 5.3: **OOD detection performance metrics using CLIP models and label names for ImageNet-1K as in-distribution**. We report the best metric in bold across different ViTs. The symbol † indicates our reproduction of Ming et al. [55].

benchmark, namely CIFAR10-A ($\lambda = 0$) and its smoothened version CIFAR10-AS ($\lambda > 0$). The generation of an adversarial example is shown in Fig. 5.1.

C.5 Experimental evaluations

First, we benchmark 25 publicly available pretrained models on common ImageNet and CIFAR OOD detection benchmarks, as illustrated in Fig. 5.2. Second, we compare PLP against the aforementioned baselines on ImageNet benchmarks in Table 5.3. For a fair comparison with [55], we use the CLIP ViT-B and ViT-L trained on OpenAI-400M [63], but also report results with the recently released CLIP ViT-G trained on LAION-2B [12]. In contrast to [39], we remove the Places OOD subset due to its high class overlap of 59.5% with ImageNet [6]; instead, we add IN-O and NINCO. Third, in Table 5.4, we compare linear probing on CLIP's visual representations to standard supervised fine-tuning for OOD detection. Fourth, in Table 5.5, we conduct further OOD detection evaluations with CLIP ViT-G, based on the availability of $\mathcal{D}_{\text{in}}$ class names or (few-shot) labeled images and different detection scores. In Table 5.6, we compare linear probing to fine-tuning for ImageNet-21K pretrained models. Finally, we study the robustness against the adversarially created OOD samples (CIFAR10-A, CIFAR10-AS) using CIFAR100 as $\mathcal{D}_{\text{in}}$.

C.6 Implementation details

Since probing and PLP only train a linear layer on precomputed representations, it is more scalable and significantly faster than fine-tuning while having minimal overhead compared to pseudo-MSP. We used the Adam optimizer [43] with a mini-batch size of 256 for CIFAR10 and CIFAR100 and 8192 for ImageNet and trained for 100 epochs with a weight decay of 10^{-3}. The learning rate is set to $10^{-3} \cdot (\text{mini-batch size})/256$ with a linear warm-up over the first ten epochs and cosine decay after that. All the experiments

Method	OOD Dataset									
	iNaturalist Plants		SUN		Places		Texture		NINCO	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIP ViT-B										
FT+Energy	42.60	83.95	60.87	73.12	64.39	69.65	61.16	73.18	67.57	70.36
FT+MSP	36.80	90.49	60.53	81.72	63.29	80.54	54.42	82.60	**63.70**	**80.94**
Prob.+Energy	**24.36**	**95.72**	**50.30**	**90.03**	**50.26**	**88.60**	**50.97**	**88.10**	78.31	79.23
CLIP ViT-L										
FT+Energy	30.35	88.75	46.75	79.42	54.60	74.11	53.03	76.37	55.70	77.70
FT+MSP	31.33	91.90	51.33	84.97	55.68	82.90	48.68	84.52	**56.79**	84.27
Prob.+Energy	**8.65**	**97.89**	**41.42**	**91.91**	**43.05**	**90.81**	**44.14**	**90.27**	67.05	**86.32**
CLIP ViT-H (LAION-2B)										
FT+Energy	27.99	90.07	39.77	83.27	49.66	77.84	42.37	81.52	**51.62**	78.18
FT+MSP	26.70	92.90	49.19	85.69	54.51	83.28	46.66	84.70	55.48	83.90
Prob.+Energy	**6.46**	**98.28**	**32.42**	**93.47**	**38.89**	**91.68**	**26.45**	**93.93**	55.39	**89.79**
CLIP ViT-G (LAION-2B)										
Prob.+Energy	6.29	98.27	32.21	93.27	37.16	92.02	24.63	94.31	49.35	90.4

Table 5.4: **Supervised OOD detection performance on ImageNet-1K using CLIP: linear Prob. outperforms fine-tuning on average, especially for larger-scale models, while being computationally cheaper and significantly faster to train**. The best method for each model architecture is bolded, excluding CLIP ViT-G. The reported results from fine-tuning are based on our evaluations using publicly available CLIP weights fine-tuned on ImageNet-1K from *timm* [79]. Prob. refers to linear probing and FT refers to fine-tuning.

were carried out in a single NVIDIA A100 with 40GB VRAM. Moreover, we emphasize that the standard deviation of probing and PLP is less than 0.01%, measured over 10 independent runs. To create the adversarial datasets CIFAR10-A and CIFAR10-AS, we perform 250 steps with the Adam optimizer with a learning rate of 10^{-3} on 1K OOD images. We set λ to $5 \cdot 10^3$ when applying smoothing (Eq. 5.8).

D Experimental Results

Unsupervised OOD detection. In Fig. 5.2, we initially investigate whether there is a connection between the $\mathcal{D}_{\text{in}}^{test}$ classification accuracy and unsupervised OOD detection AUROC by benchmarking 25 feature extractors. Out of them, CLIP models exhibit the strongest correlation (R^2 coefficient ≥ 0.92) independent of their network architecture (i.e. ConvNext, ViTs, etc). CLIP's best instances (ViT-G, ConvNext-XXL) are currently the best-performing unsupervised OOD detectors, aligning with recent results from [22]. The features of CLIP ViT-G even outperform the ones from supervised training on ImageNet as $\mathcal{D}_{\text{in}}$. Moreover, we observe that when the $\mathcal{D}_{\text{pretrain}}$ is different from $\mathcal{D}_{\text{in}}$ (CIFAR benchmarks), a positive correlation can still be identified for self-supervised and supervised models pretrained on ImageNet ($R^2 = 0.82$), yet not as strong as CLIP ($R^2 = 0.95$).

Large-scale OOD detection with access to class names. We report absolute gains and AUROC scores. In Table 5.3, we compare pseudo-MSP, based on our reproduction of [55], to PLP on all five large-scale benchmarks and find an average improvement of 2.81%, 1.85%, 3.44% AUROC for CLIP ViT-L, ViT-H, and ViT-G, respectively. The reported gains are computed using the energy score, which is slightly superior to RMD on average and significantly faster to compute. Interestingly, PLP has consistent improve-

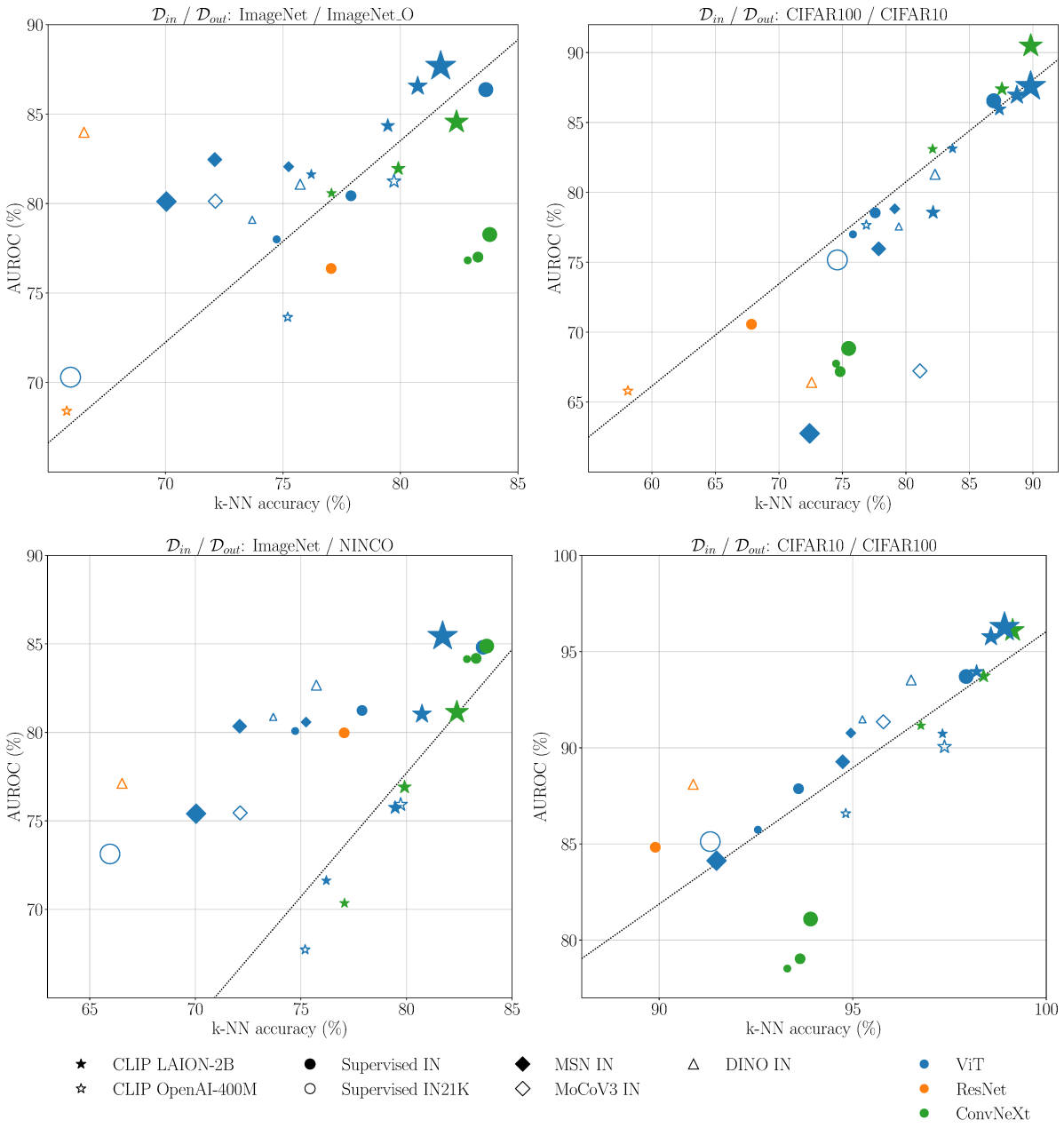


Figure 5.2: **In-distribution test accuracy using k=20 nearest neighbours (k-NN) (x-axis) versus unsupervised out-of-distribution (OOD) detection score (AUROC %) (y-axis) for ImageNet→ImageNet-O, CIFAR100→CIFAR10, ImageNet→NINCO and CIFAR10→CIFAR100** [16, 37, 44]. CLIP models [63, 12] exhibit a strong dependence between OOD performance and in-distribution test accuracy even without fine-tuning. The black dotted line is fitted using the CLIP models (starred datapoints), where a coefficient $R^2 > 0.92$ for all benchmarks was found. The OOD detection score is computed using the top-1 cosine similarity. Different colors are utilized for different architectures, while symbol sizes indicate architecture size. IN indicates ImageNet and IN21K indicates ImageNet-21K.

CLIP ViT-G/14	$\mathcal{D}_{\text{in}}$ labels/ names	CIFAR100 CIFAR10	CIFAR10 CIFAR100	IN IN-O	IN NINCO
k-means MD	✗/ ✗	72.8	89.5	87.6	80.7
1-NN	✗/ ✗	**87.6**	**98.2**	**88.0**	**84.0**
Pseudo-MSP	✗/ ✓	94.2	97.3	88.1	83.4
PLP + MSP	✗/ ✓	92.7	97.9	86.6	86.7
PLP + RMD	✗/ ✓	**97.1**	98.3	**91.9**	88.4
PLP + Energy	✗/ ✓	95.0	**98.5**	91.4	**88.7**
MD	✓/ ✓	73.1	91.1	88.1	81.5
RMD	✓/ ✓	96.3	98.8	92.4	89.3
FS_{10}+MSP	✓/ ✓	89.4	96.5	88.1	84.8
FS_{10}+Energy	✓/ ✓	90.9	96.6	90.8	83.0
Probing+MSP	✓/ ✓	94.1	98.7	88.1	89.1
Probing+RMD	✓/ ✓	**97.3**	98.8	92.5	89.5
Probing+Energy	✓/ ✓	96.3	**99.1**	**92.9**	**90.4**

Table 5.5: **OOD detection AUROCs (%) for multiple evaluations and scores using CLIP ViT-G/14 trained on LAION-2B.** IN stands for ImageNet-1K and FS_{10} for few-shot fine-tuning with $p = 10$.

ments on the newly proposed NINCO dataset, wherein all samples have been visually verified to be semantically different from $\mathcal{D}_{\text{in}}$, with an average AUROC gain of 3.93% across CLIP models. We highlight that the discrepancy between our reproduction and the reported results of [55] using ViT-B and ViT-L is *not* due to different CLIP weights.

Large-scale supervised OOD detection. As shown in Table 5.4, fine-tuning CLIP models is often unnecessary for supervised OOD detection. More precisely, probing on larger models such as CLIP ViT-H consistently outperforms fine-tuning by a large margin of 7.34% on average. The performance discrepancy of CLIP ViT-L on Places and SUN compared to the other benchmarks is partially attributed to the class overlap. By contrast, iNaturalist, Texture, and NINCO are sufficiently disjoint from ImageNet in terms of $\mathcal{D}_{\text{in}}$ class overlap.

Adversarial OOD detection robustness. By evaluating CLIP ViT-G on the introduced CIFAR100→CIFAR10-A benchmark, we found that it is possible to drop the AUROC score from 87.6% →50.3% using 1-NN and 94.2%→49.5% using pseudo-MSP. Introducing the smoothness restriction (Eq. (5.8)) degrades performance to 55.8% and 51.9% AUROC on CIFAR100→CIFAR10-AS using 1-NN and pseudo-MSP, respectively. Note that an AUROC score of 50% is a random guess's score, meaning that CLIP ViT-G performs slightly better than a random guess.

Ablation study for CLIP ViT-G for all OOD detection setups. In Table 5.5, we conduct additional experimental evaluations using CLIP ViT-G for all three OOD detection scenarios (unsupervised, class names are available, supervised). In the unsupervised case, k-means + MD yields an inferior AUROC compared to 1-NN, precisely lower by 6.8% on average. By incorporating the $\mathcal{D}_{\text{in}}$ class names using CLIP's text encoder, we find that the PLP consistently outperforms pseudo-MSP with a mean improvement of 3.16% and 2.63% over PLP+RMD and PLP+Energy respectively. In the supervised sce-

	Finetuned on $\mathcal{D}_{\text{in}}$	IN1K NINCO	IN1K IN-O	CIFAR100 CIFAR10	CIFAR10 CIFAR100
ConvNext-B	✗	**95.4**	**94.9**	**93.0**	**98.0**
ConvNext-B	✓	88.8	85.8	90.8*	96.9*
ViT-L	✗	**95.2**	**95.3**	89.8	94.3
ViT-L	✓	91.3	92.1	**97.9***	**98.5***
ResNet50+ViT-B	✗	**95.9**	**95.8**	89.2	96.6
ResNet50+ViT-B	✓	92.9	92.1	**96.2***	**98.5***

Table 5.6: **Comparing the OOD detection AUROC (%) of ImageNet-21K pretrained models: linear probing versus fine-tuning.** We use publicly available models with and without fine-tuning on ImageNet from *timm* and evaluate them using the energy score. Following [21], we use the MD on CIFAR fine-tuned models (indicated with an asterisk), which is inefficient at the scale of ImageNet.

nario, we highlight that RMD is a strong baseline, surpassing MD by 10.75% AUROC on average, while our PLP+Energy marginally improves RMD by 0.47%. MSP is constantly the worst choice as an OOD detection score after linear probing.

E Discussion

Pixel-related features in CLIP's learned representations. Similar to [54] for image classification, the OOD detection performance close to a random guess demonstrates that even the top-performing CLIP models trained on billion-scale image-text pairs can be fooled by a visual signal manipulation that is invisible to humans. This finding gives further evidence that besides label-related features, (local) pixel information affects the learned representations and needs to be explored in greater depth [61, 18]. Another possible research avenue is how adversarial OOD samples transfer between different feature extractors, which is left for future work.

Does PLP improve $\mathcal{D}_{\text{in}}^{\text{test}}$ accuracy? By comparing the $\mathcal{D}_{\text{in}}^{\text{test}}$ accuracy on ImageNet of the trained head with the pseudo-labels versus the zero-shot classification of CLIP ViT-G, we found a minor accuracy improvement of 0.5% on ImageNet. However, accuracies deteriorate for CLIP ViT-B and CLIP ViT-L on average, which suggests that their $\mathcal{D}_{\text{in}}$ text-based pseudo-labels have less true positives. The fact that PLP improves the OOD detection performance (compared to pseudo-MSP) without significantly increasing $\mathcal{D}_{\text{in}}^{\text{test}}$ accuracy is a new direction left for future work. Another interesting avenue is whether the text-based pseudo-labels from CLIP can be used with other, possibly smaller-scale, visual feature extractors.

Does linear probing lead to similar OOD detection performance compared to fine-tuning for ImageNet-21K pretrained models? In Table 5.6, we compare linear probing versus fine-tuning for 3 publicly available models, where there is a corresponding fine-tuned model on ImageNet. Even though fine-tuning is the standard practice in supervised OOD detection [21], the obtained results indicate that it performs inferior to linear probing on the ImageNet OOD-related benchmarks even for ImageNet-21K pretrained models. Fine-tuning is, however, still the best approach on the small-scale

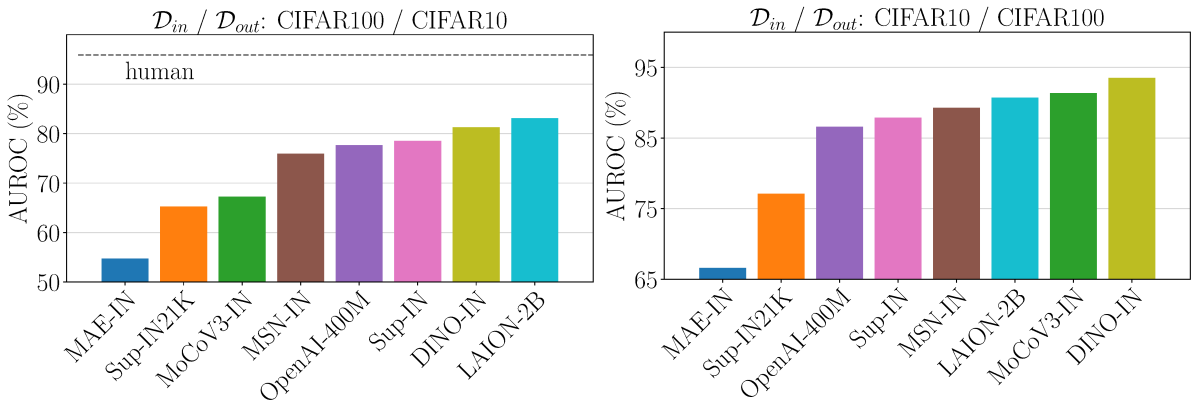


Figure 5.3: **AUROC values for unsupervised OOD detection using ViT-B/16 pretrained on different datasets (IN, IN-21K, OpenAI-400M, LAION-2B) and pretext tasks**. IN indicates ImageNet. The performance of CLIP on LAION-2B is bottlenecked by the model size as reported in [12] and larger networks are needed when scaling up to billion-scale datasets. The horizontal line indicates human-level performance, as reported in Fort et al. [21].

OOD benchmarks (e.g. CIFAR10 → CIFAR100), given ImageNet21K pretraining. This explains why a simple and scalable approach such as probing may have been overlooked. As suggested by [39], we confirm that OOD detection approaches tested on small-scale benchmarks do not always translate effectively into larger-scale setups, where simpler methods need to be revisited. Combined with the results from Table 5.4, we validate that the OOD-related information is readily available on large-scale foundational models, as recently stated in [40, 60].

Is the choice of the best feature extractor consistent for smaller-sized models? Not in the scale of ViT-B and CIFAR benchmarks. In Fig. 5.3, we keep the model architecture fixed (ViT-B) and visualize the unsupervised OOD detection performance (1-NN) for different datasets (ImageNet, ImageNet21K, OpenAI-400M, LAION-2B) and pretraining types (CLIP, supervised, self-supervised). The rationale behind using the CIFAR-based benchmarks is to focus on feature transferability. We highlight that large pretraining datasets such as LAION-2B are bottlenecked by the model size of ViT-B as explained in [12]. Unlike Fig. 5.2, we show that the ranking of backbones is not always consistent between different $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$ at this scale. Finally, DINO ViT-B surpasses supervised ImageNet pretraining on both benchmarks, outlining the transferability of features of self-supervised methods, which is consistent with the results of [19, 58, 1].

F Conclusion

This work presented a thorough experimental study by leveraging pretrained models for visual OOD detection, focusing on CLIP. It was demonstrated that CLIP models are powerful unsupervised OOD detectors, outperforming even $\mathcal{D}_{\text{in}}$ supervised models on large-scale OOD detection benchmarks. A fast, simple, OOD data-agnostic, and scalable

method called PLP that trains a linear layer based on $\mathcal{D}_{\text{in}}$ text-based pseudo-labels was introduced. PLP outperformed the previous state-of-the-art (pseudo-MSP) for most CLIP models and ImageNet-based benchmarks while substantially improving the larger architectures (i.e. ViT-G achieves a mean AUROC gain of 3.4%). Furthermore, it was demonstrated that probing can replace the costly fine-tuning on CLIP and even ImageNet-21K models on large-scale benchmarks, where CLIP ViT-H exhibited an average AUROC improvement of 7.3%. Finally, a simple feature-based adversarial OOD data manipulation method showed that billion-scale feature extractors (CLIP ViT-G) are vulnerable to adversarial OOD attacks.

References

[1] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. *arXiv preprint arXiv:2303.17896*, 2023.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 456–473. Springer, 2022.

[4] Mohammad Azizmalayeri, Arshia Soltani Moakar, Arman Zarei, Reihaneh Zohrabi, Mohammad Taghi Manzuri, and Mohammad Hossein Rohban. Your out-of-distribution detection method is not robust! In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[5] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023.

[6] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[9] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. *arXiv preprint arXiv:2003.09711*, 2020.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[12] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

[13] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[14] Niv Cohen, Ron Abutbul, and Yedid Hoshen. Out-of-distribution detection without class labels. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 101–117. Springer, 2023.

[15] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[18] Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1934–1943, 2023.

[19] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[20] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[21] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

[22] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[24] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[25] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[26] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018.

[27] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[28] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 2022.

[31] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[32] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[33] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[34] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *ACL*, pages 2744–2751, 2020.

[35] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[36] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[37] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[38] Wen Heng, Shuchang Zhou, and Tingting Jiang. Harmonic adversarial attack method. *arXiv preprint arXiv:1807.10590*, 2018.

[39] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[40] Nathan Inkawhich, Gwendolyn McDonald, and Ryan Luley. Adversarial attacks on foundational vision models. *arXiv preprint arXiv:2308.14597*, 2023.

[41] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[42] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[43] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[44] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[45] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[46] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[47] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[49] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[50] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Klaus Robert Muller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*, 2022.

[51] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[52] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2022.

[53] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[54] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*, 2023.

[55] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022.

[56] Sina Mohseni, Arash Vahdat, and Jay Yadawa. Shifting transformation learning for out-of-distribution detection. *arXiv preprint arXiv:2106.03899*, 2021.

[57] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

[58] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[59] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[61] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2023.

[62] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[64] Nima Rafiee, Rahil Gholamipoor, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, and Markus Kollmann. Self-supervised anomaly detection by self-distillation and negative sampling. In *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings; Part IV*, pages 459–470. Springer, 2022.

[65] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021.

[66] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[67] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

[68] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[69] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[70] Vikash Sehwag, Mung Chiang, and Prateek Mittal. {SSD}: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021.

[71] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[72] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[73] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[74] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[75] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.

[76] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[77] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

[78] Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *arXiv preprint arXiv:2207.09239*, 2022.

[79] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[80] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[81] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[82] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[83] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[84] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[85] Heng Yin, Hengwei Zhang, Jindong Wang, and Ruiyu Dou. Boosting adversarial attacks on neural networks with better optimizer. *Secur. Commun. Networks*, 2021:9983309:1–9983309:9, 2021.

[86] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[87] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 102–117. Springer, 2020.

[88] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Appendix

App. I Text context prompts and weights.

Following [55], for the text encoder of CLIP, we take the mean of the L2-normalized text representations of the following prompts: "an image of a {label}", "a photo of a {label}", "a blurry photo of a {label}", "a photo of many {label}", "a photo of the large {label}", "a photo of the small {label}". We randomly select $p = 10$ images per class for few-shot probing and take the average AUROC over 5 runs. We set the mini-batch size to 32 for CIFAR100 and 1024 for ImageNet. To enforce reproducibility for our results on Table 6, the corresponding models can be found using the *timm* (version 0.6.12) names [79]: *vit_large_patch16_224_in21k*, *convnext_base_in22k*, and *vit_base_r50_s16_224_in21k*. We note that probing takes less than 15 minutes on ImageNet on a single GPU.

App. II Computational complexity of OOD detection methods.

As reported in Table 6, [21] used the MD as an OOD score, which is inefficient at the scale of ImageNet and can become prohibitively slow for even larger datasets. Given a feature dimension d and dataset size N, the total time complexity scales linearly with dataset size (for the computation of the covariance matrix) and in cubic time with d due to the inverse calculation of the covariance matrix, resulting in $O(Nd\min(N, d) + d^3) = O(Nd^2 + d^3)$ since $N > d$. We thus conclude that, in addition to the performance metrics, the computational complexities need to be taken into account in future studies to design scalable and efficient OOD detection systems.

App. III Few-shot evaluations on CIFAR100→ CIFAR10.

In Fig. 5.4, we show that even with 1% of the data, we can surpass the zero-shot 1-NN score. RMD applied on the logits seems to be consistently better than MSP.

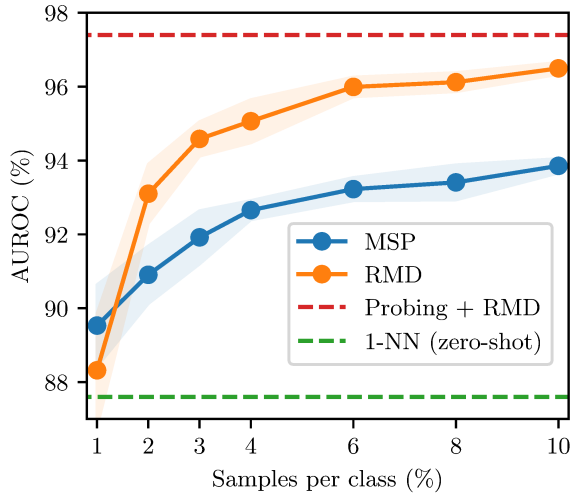


Figure 5.4: **Few-shot linear probing on CIFAR100 → CIFAR10.** Samples per class (shown as a percentage %) versus OOD detection performance (*y-axis*).

App. IV Adversarial examples.

We illustrate more adversarially generated samples using the proposed method in Fig. 5.5. The created adversarial OOD datasets CIFAR10-A and CIFAR10-AS are publicly available via this hyperlink.

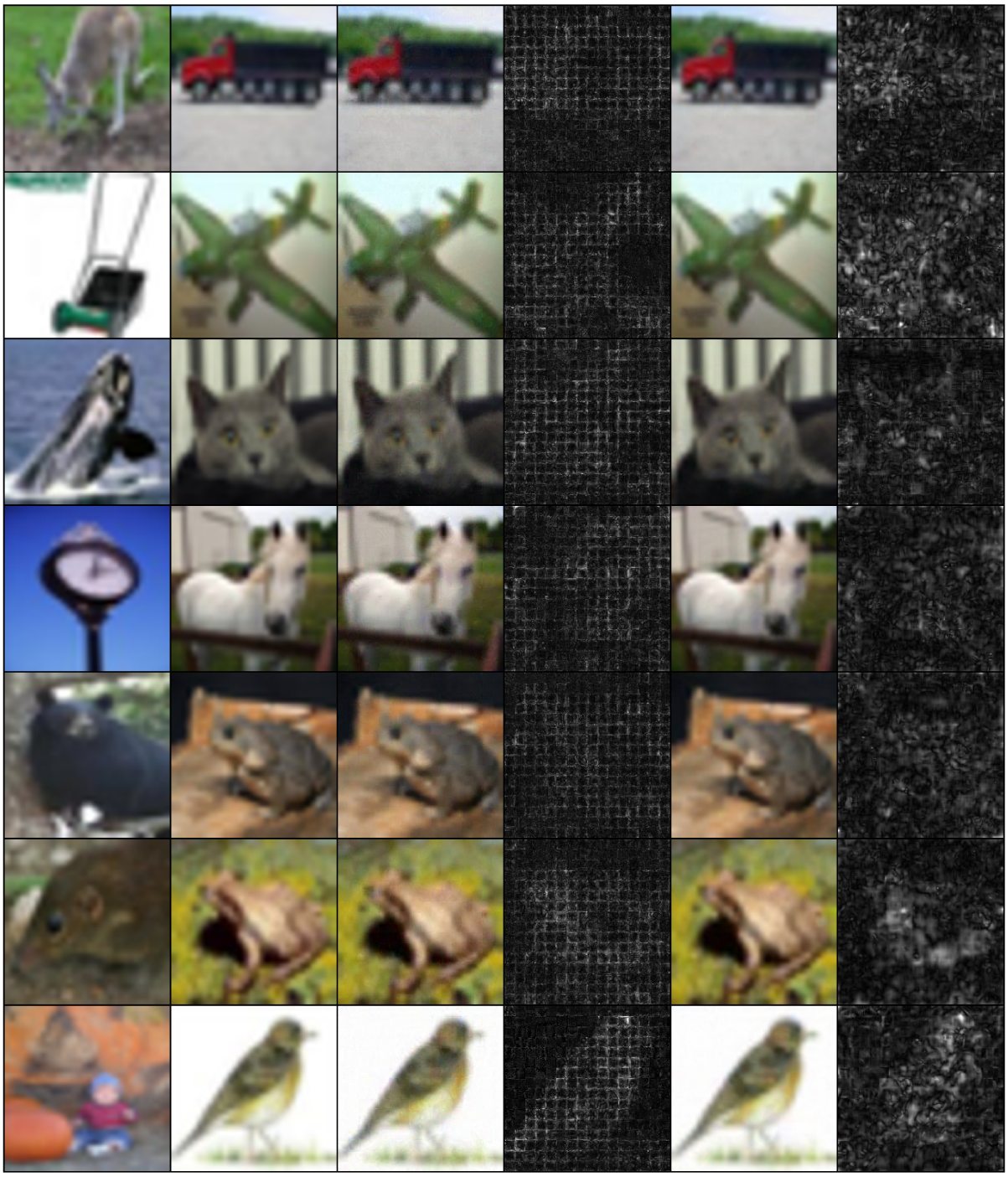


Figure 5.5: **Additional CIFAR10 adversarially manipulated examples for CIFAR100 → CIFAR10 OOD detection with (CIFAR10-AS) and without (CIFAR10-A) the smoothing constraint.** Columns from *left to right*: target in-distribution image from CIFAR100, original CIFAR10 sample, adversarially manipulated image without smoothing, the Euclidean pixel-wise distance between the original image and perturbed image, adversarial example with smoothing, Euclidean distance.

6. Rethinking Cluster-Conditioned Diffusion Models for Label-Free Image Synthesis

Outline

A Introduction 1
B Related Work 2
C Method 3
D Experimental Evaluation 6
E Discussion 10
F Conclusion 15
References 16
Appendix 22

Bibliographic Information

Adaloglou, N., Kaiser, T., Michels, F., & Kollmann, M. (2025, February). Rethinking cluster-conditioned diffusion models for label-free image synthesis. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 3603-3613). IEEE.

Corresponding author: *adaloglo@hhu.de*

Copyright Notice

This chapter is based on the author's paper published in WACV 2025, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2025. The author retains the copyright to this work. The accepted version of this article is available in this link. The copyright was adjusted according to the guidelines of the publisher.

Acknowledgements

The author would like to thank Tim Kaiser for sharing his expertise regarding training and sampling from the diffusion model baseline, which significantly accelerated the project.

Individual Contributions

The author identified several prior works that employed alternative forms of generative conditioning (i.e. segmentation masks). Building upon discussions with his supervisor, the author proposed leveraging clusters derived from learned feature representations. The initial implementation of this approach was carried out jointly by Tim Kaiser and the author, yielding promising results on the CIFAR10 dataset. Beyond demonstrating superior performance, the author discovered that cluster-based conditioning offers substantial training sample efficiency relative to unconditional generation. After the preliminary results, the author conducted a comprehensive investigation into multiple complementary aspects. Such aspects include, but are not limited to, the selection of the number of clusters, comparison of grouping strategies, qualitative visual analyses, and the empirical relationship between the downstream task of image clustering and cluster-conditional image synthesis. Finally, the author served as the lead writer of the manuscript.

Abstract

Diffusion-based image generation models can enhance image quality when conditioned on ground truth labels. Here, we conduct a comprehensive experimental study on image-level conditioning for diffusion models using cluster assignments. We investigate how individual clustering determinants, such as the number of clusters and the clustering method, impact image synthesis across three different datasets. Given the optimal number of clusters with respect to image synthesis, we show that cluster-conditioning can achieve state-of-the-art performance, with an FID of 1.67 for CIFAR10 and 2.17 for CIFAR100, along with a strong increase in training sample efficiency. We further propose a novel empirical method to estimate an upper bound for the optimal number of clusters. Unlike existing approaches, we find no significant association between clustering performance and the corresponding cluster-conditional FID scores. Code is available at `https://github.com/HHU-MMBS/cedm-official-wavc2025`

A Introduction

Diffusion models have enabled significant progress in many visual generative tasks, such as image synthesis [24, 34, 52] and manipulation [74]. Conditioning diffusion models on human-annotated data is today's standard practice as it significantly improves the image fidelity [19, 6, 20]. Image-level conditioning is typically realized by using associated text captions [60, 61] or class labels, if available [24, 48]. Nonetheless, human-annotated labels are costly and often contain inaccuracies [8, 5], while publicly available image-text pairs can be non-descriptive [7].

Large, human-annotated datasets are prohibitively costly [54] and hard to obtain in a plethora of real-life applications such as medical image synthesis [36]. Human annotation is subject to the label collection procedure and can have varying annotation hierarchies [26, 8, 3]. An ideal conditioning signal would be based solely on shared characteristics corresponding to general visual concepts, as illustrated in Fig. 6.1. To provide a ground truth (GT) label-free alternative that is applicable to unlabelled datasets, we focus on cluster-based conditioning on diffusion models. Image clustering refers to algorithmically assigning a semantic group to an image, called a cluster, given an *a priori* number of such groups [11, 15, 33].

Maximizing cluster alignment with ground truth (GT) labels does not guarantee optimal cluster-conditional image generation. To create visually distinguishable groups (Fig. 6.1), a clustering algorithm would sometimes group images into more fine-grained groups than the GT labels, i.e., sub-grouping analog and digital devices that map to the same label or merge similarly looking images with different labels (merging different types of trees).

Deep image clustering has recently seen significant progress [72], especially when utilizing off-the-shelf features learned from large-scale pre-training [70, 3, 3]. Currently, the relationship between clustering performance and cluster-conditional image synthesis performance remains unclear for several reasons. First, clustering metrics capture the alignment to GT labels, whereas generative models may benefit more from visually distinguishable groups [29]. Second, clustering metrics cannot be computed for unlabeled datasets. Third, existing metrics are not suitable for fairly evaluating cluster assignments with varying numbers of clusters. Therefore, they cannot help to determine the optimal cluster granularity w.r.t. generative performance for an arbitrary dataset. This aspect has not been sufficiently explored to date [45, 20].

Existing cluster-conditioned generative approaches [45, 20, 29] often adopt simple generative baselines [24] and have mainly investigated k-means clustering on balanced classification datasets. The adoption of k-means is justified due to its ability to work with off-the-shelf feature extractors and its non-parametric nature, apart from setting the number of clusters. Conversely, it is well-established that k-means is suboptimal for image clustering because its cluster assignments are highly imbalanced [70, 29]. While cluster-conditioned approaches are sensitive to the number of clusters [20], there is currently no method to predetermine this number. Previous methods have overlooked this [45, 20, 29], significantly hindering their applicability to unlabeled datasets. Recent feature-based image clustering methods have demonstrated superior performance compared to k-means [70, 3], and their efficacy in conditional image generation is yet to be investigated. For the above reasons, cluster-conditioning methods underperform GT conditional ones to

Figure 6.1: An ideal image-level conditioning should group images based on shared patterns, shown in the same row, which do not always align with human labels, indicated above each image (CIFAR100 [40] samples).

date.

In this paper, we systematically study clustering as a label-free condition for diffusion models. We demonstrate that optimal cluster granularity achieves state-of-the-art FID scores (1.67 for CIFAR10 and 2.17 for CIFAR100) while enhancing training sample efficiency. We propose a computationally efficient method to derive an upper cluster bound using feature-based clustering, which narrows the search space for optimal image synthesis performance. We validate this upper bound across three datasets and two generative metrics. Finally, we find no significant association between clustering performance and cluster-conditional image synthesis performance across various clustering methods, performance metrics, and pre-trained models.

B Related Work

B.1 Conditional generative models

Early attempts on controllable image synthesis adopted generative adversarial networks (GANs) [47]. By conditioning both the generator and discriminator on human labels, GANs can produce images for a specific GT label, even at the scale of ImageNet [10]. Recently, diffusion models (DMs) have emerged as an expressive and flexible category of generative models [63, 64, 25, 58, 35]. Internal guidance-based methods [19, 25, 27, 28] further improved the flexibility and visual fidelity of DMs during sampling such as classifier and classifier-free guidance. DMs exhibit an enormous number of label-conditional variants [24, 39, 34, 48, 53]. Recently, Stein et al. [65] proposed the adoption of Fréchet DINOv2 [49] distance as it aligns better with human preferences and demonstrated that DMs achieve the highest perceptual realism.

B.2 Alternative conditioning of generative models

Unlike internal guidance [27, 25, 28], external conditioning signals are computed from the training set using additional models, which is commonly referred to as *self-conditioning*. Self-conditioning signals can be roughly divided into image-level and sub-image-level. Image-level conditioning refers to a single condition for all pixels in an image, such as cluster assignments or text captions [58]. Sub-image-level conditions refer to specific parts or regions of an image [29, 4]. For instance, Hu et al. [29] extend DMs to incorporate conditioning of bounding boxes or segmentation masks.

Feature-based conditioning. Using image or text features can also provide an informative conditioning signal [9, 62, 44, 56]. Bordes et al. [9] condition DMs directly on image representations (*oracle features from real images*) and reveal their impact on visual fidelity. Instance-conditioned GAN combines GT labels with features from each image's NN set [13]. Ramesh et al. [56] leverage the zero-shot capability of CLIP to condition DMs on language-guided image representations to improve diversity. Zhou et al. provide a language-free framework [76] that directly generates text features for images using CLIP. Still, CLIP models require an *a priori* set of descriptive candidate captions, such as label names.

Cluster-based conditioning. Similar to label conditioning, conditioning on cluster assignments facilitates DMs by allowing them to specialize on a distinct set of shared visual features [29, 20, 45]. In [20], Bao et al. computed the k-means clusters offline using contrastive learned features on the training data. However, their approach could not outperform the label-conditioned models. The first cluster-based approach that achieves competitive performance compared to GT labels leverages pre-trained feature extractors [29], which is the closest to our work. In [29], the authors attempt to weak establish a correlation between clustering and cluster-conditional generative performance for different self-supervised models. The observed correlation cannot be tested on unlabeled datasets, and its sensitivity to the number of clusters is unknown. While it was demonstrated that DINO [12] provides the most informative k-means clusters [29], their method lacks a strategy for choosing the cluster granularity for unlabeled datasets.

C Method

C.1 Notations and prerequisites

We consider the frequent scenario where we have access to an **unlabeled** dataset D and a pre-trained feature extractor $g(\cdot)$. We denote the number of ground truth (GT) labels as C_{GT}, which we assume to be unknown. To distinguish cluster-based conditioning from the image clustering task, we denote the number of *visual groups* C_V of D as the optimal number of clusters w.r.t. image synthesis (e.g. as measured by FID [23]). We adopt the diffusion model approach of Karras et al. (EDM [34]) as a baseline, which introduced various improvements to the standard DDPM [24]. Although we focus on EDM throughout this work, our method can be applied to any conditional generative model.

TEMI clustering. Given an *a priori* determined number of clusters C and a feature extractor $g(.)$, TEMI [2] first mines the m nearest neighbors (NN) of all $\mathbf{x} \in D$ in the

feature space of $g(\cdot)$ based on their cosine similarity. We denote the set of NN for $\mathbf{x}$ by $S_{\mathbf{x}}$. During training, TEMI randomly samples $\mathbf{x}$ from D and $\mathbf{x}'$ from $S_{\mathbf{x}}$ to generate image pairs with similar (visual) features. A self-distillation framework is introduced to learn the cluster assignments with a teacher and student head $h_t(\cdot)$ and $h_s(\cdot)$ that share the same architecture (i.e. 3-layer MLP) but differ w.r.t. their parameters. The features of the image pair are fed to the student and teacher heads $h_s^i(\mathbf{z})$, $h_t^i(\mathbf{z}')$, where $i \in \{1, \ldots, H\}$ is the head index and $\mathbf{z} = g(\mathbf{x})$, $\mathbf{z}' = g(\mathbf{x}')$. The outputs of the heads are converted to probabilities $q_s(c|\mathbf{x})$ and $q_t(c|\mathbf{x}')$ using a softmax function. The TEMI objective,

$$\mathcal{L}^i_{\text{TEMI}}(\mathbf{x}, \mathbf{x}') := -\frac{1}{H}\sum_{j=1}^{H}\sum_{c'=1}^{C} q_t^j(c'|\mathbf{x}) q_t^j(c'|\mathbf{x}') \\ \log \sum_{c=1}^{C} \frac{(q_s^i(c|\mathbf{x}) q_t^i(c|\mathbf{x}'))^\gamma}{\tilde{q}_t^i(c)} \quad , \tag{6.1}$$

maximizes the pointwise mutual information between images $\mathbf{x}$ and $\mathbf{x}'$, using the clustering index c as information bottleneck [33]. Here, $\tilde{q_t}^i(c)$ is an estimate of the teacher cluster distribution $\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[q_t^i(c|\mathbf{x})]$, which can be computed by an exponential moving average over mini-batches B defined as $\tilde{q_t}^i(c) \leftarrow \lambda\, \tilde{q}_t(c) + (1-\lambda)\frac{1}{|B|}\sum_{\mathbf{x}\in B} q_t^i(c|\mathbf{x})$.

The hyperparameter $\gamma \in (0.5, 1]$ avoids the collapse of all sample pairs into a single cluster, and $\lambda \in (0, 1)$ is a momentum hyperparameter. The loss function Eq. (6.1) is further symmetrized and averaged over heads to obtain the final training loss. After training, the cluster assignments $c^*(\mathbf{x}) = \arg\max_c q_t(c|\mathbf{x})$ and the empirical cluster distribution $q(c)$ on the training set is computed from the TEMI head with the lowest loss. We use $c^*(\mathbf{x})$ as a condition of the generative model during training and sample from $q(c)$ to generate images. Adaloglou et al. [2] show that the value $\gamma = 0.6$ enforces a close to uniform cluster utilization and achieves state-of-the-art clustering accuracy for C_{GT}[2].

C.2 Cluster-conditional EDM (C-EDM)

We consider k-means and TEMI clusters using off-the-shelf feature extractors with C_V clusters to compute cluster assignments. We experimentally observed that $C_V > C_{GT}$, even though this is not a requirement of our approach. We highlight that clustering using $C > C_{GT}$ while enforcing a uniform cluster utilization (such as TEMI with $\gamma = 0.6$) has not been previously explored [2, 43, 22, 70] as it reduces the cluster alignment with the GT labels. We denote the cluster-conditioned EDM model as *C-EDM*.

A direct estimation of C_V is hard to obtain for unlabeled datasets. In principle, C_V can be found using a hyperparameter search with FID as an evaluation metric, which is computationally expensive. In contrast to existing approaches that perform a restricted hyperparameter search around C_{GT}, we propose a new metric that allows us to derive an upper bound for C_V (Section C.3), which requires no prior knowledge about the dataset.

Generative methods	CIFAR10		CIFAR100		FFHQ-64	FFHQ-128
	Unlabelled	GT	Unlabelled	GT	Unlabelled	Unlabelled
DDPM [24, 29]	3.17	-	10.7	9.7	-	-
PFGM++ [73]	-	1.74	-	-	2.43	-
EDM [34]	1.97	1.79	-	-	2.39	-
EDM (our reproduction)	2.07	1.81	3.41	2.21	2.53	5.93
Self-conditional methods						
IDDPM+k-means (k=10) [20]	2.23	-	-	-	-	-
DDPM+k-means† (k=400) [29]	-	-	9.6	-	-	-
C-EDM+k-means† (Ours)	1.69	-	2.21	-	**1.99**	-
C-EDM+TEMI† (Ours)	**1.67**	-	**2.17**	-	2.09	**4.40**
EDM+oracle features† (Ours)	2.21	-	2.25	-	1.77	-

Table 6.1: State-of-the-art generative model comparison: FID (↓) for various GT labeled and unlabelled benchmarks. We use C_V =100, 200, 400 clusters for CIFAR10, CIFAR100, and FFHQ-64, respectively. Self-conditional methods with † use the pre-trained DINO ViT-B feature extractor. Ground truth and oracle feature conditioning results are marked in gray color as they are **not** a fair comparison.

C.3 Estimating the upper cluster bound

We aim to find an upper cluster bound C_{max} using TEMI ($\gamma = 0.6$), such that $C_V < C_{max}$. During this computation, we neither use the generative model nor any additional information. We highlight that previous works [45, 20, 29] have overlooked this design choice and iterate within a small range around C_{GT}. We denote the number of utilized clusters (one training sample is assigned to it) as C^u after TEMI clustering. The TEMI cluster utilization ratio is defined as $r_C := \frac{C^u}{C} \leq 1$. Importantly, there is no guarantee that the full cluster spectrum will be utilized. Unlike k-means, which always has $r_C = 1$, we observe that as the number of clusters C increases, the TEMI cluster utilization ratio r_C typically decreases, which provides dataset-specific information. Intuitively, TEMI clustering with $\gamma = 0.6$ is enforcing $r_C \to 1$, and the observation of $r_C < 1$ is an indication that the maximum number of TEMI clusters is reached for $\gamma = 0.6$.

The search for C_{max} involves doubling the number of clusters C at each iteration until the r_C falls below a threshold, $r_C \leq \alpha$, followed by a more fine-grained grid search. C_{max} is then defined as the highest C for which $r_C > \alpha$. This results in a worst-case time complexity of $O(\log_2(\frac{C_{max}}{C_{\text{start}}}))$. This is conceptually similar to the elbow method [69, 37]. In this sense, the proposed heuristic has the same limitations as the elbow method. Yet, it is currently the only method that provides a practitioner with a starting point for an unlabelled dataset. After detecting C_{max}, we perform a bounded grid search to find $C_V \in [2, C_{max})$ using the generative model. To obtain a simple cross-dataset estimate, we empirically find a cutoff threshold $r_C \leq 0.96$ to work well across three datasets.

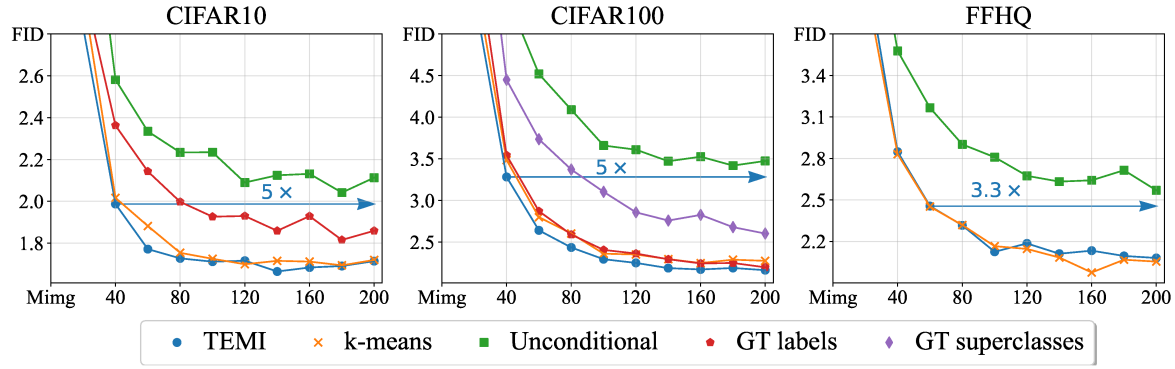


Figure 6.2: FID (*y-axis*) versus seen samples during training in millions (*x-axis*). TEMI and k-means clusters are computed using the representations of DINO ViT-B. We used $C_V = 100, 200, 400$ for CIFAR10, CIFAR100 and FFHQ-64 respectively. The training sample efficiency compared to the unconditional baseline is indicated by the arrow. Best viewed in color.

D Experimental Evaluation

D.1 Datasets, models, and metrics

Following prior works [29, 3], we use DINO ViT-B pre-trained on ImageNet [18] for image clustering. We report the Fréchet inception distance (FID) [23] to quantify the image generation performance as it simultaneously captures visual fidelity and diversity. To facilitate future comparisons, we also compute the Fréchet DINOv2 distance (FDD) [65] that replaces the features of InceptionNet-v3 [67, 68] with DINOv2 ViT-L [49]. FID and FDD are averaged over three independently generated sample sets of 50K images each. To measure the cluster alignment w.r.t. GT labels, we use the adjusted normalized mutual information (ANMI) as in [30].

We follow the default hyperparameter setup for EDM [34] and TEMI [2]. We denote the number of samples (in millions) seen during training as M_{img}. We use $M_{img} = 200$ when comparing C-EDM with other baselines and state-of-the-art methods (Table 6.1), and $M_{img} = 100$ when comparing across different number of clusters. We train the TEMI clustering heads for 200 epochs per dataset. All the experiments were conducted on 4 NVIDIA A100 GPUs with 40GB VRAM each. On this hardware, TEMI clustering was more than 50× faster than training EDM on FFHQ-64, which requires more than 2 days for 200 M_{img} with a batch size of 512. We verify our approach across CIFAR10, CIFAR100 [40] and FFHQ. CIFAR10 and CIFAR100 have 50K samples and 32^2 resolution images, while we use the 64x64 and 128x128 versions of FFHQ (FFHQ-64, FFHQ-128) consisting of 70K samples. Finally, diffusion sampling methods are not included, as they can be incorporated into any diffusion model [38, 59, 25].

D.2 State-of-the-art comparison for image synthesis

Comparison with state-of-the-art unconditional generative models. With respect to FID, we found $C_V = 100, 200, 400$ to be close to optimal cluster granularities

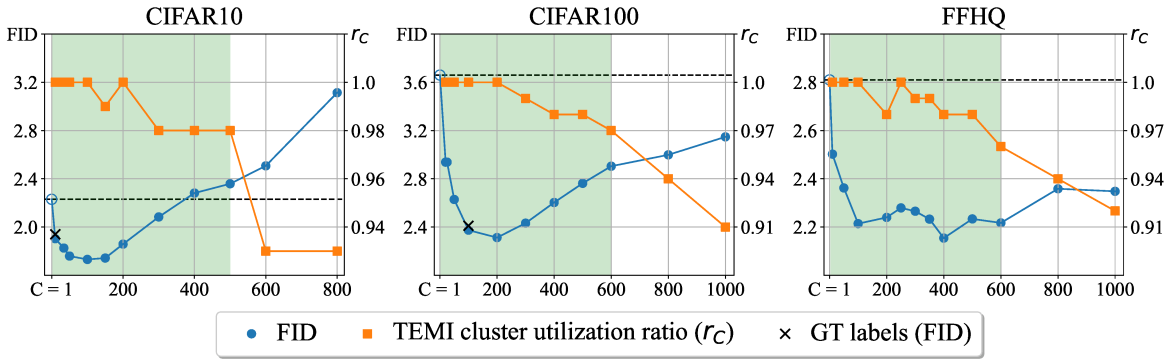


Figure 6.3: FID (*left y-axis*) and TEMI cluster utilization ratio r_C (*right y-axis*) across different numbers of clusters C (*x-axis*) using C-EDM, evaluated at $M_{img} = 100$. The green area indicates the discovered cluster range $[2, C_{max})$ for $r_C \leq \alpha = 0.96$.

for CIFAR10, CIFAR100, and FFHQ-64, respectively. We use these values to report FIDs compared to GT conditional and unconditional generative models in Table 6.1. Compared to unconditional methods, C-EDM achieves an average relative FID improvement of 24.4% and 24.9% using TEMI and k-means clusters, respectively. More importantly, previous cluster-conditional approaches did not achieve near-state-of-the-art FIDs because they: a) adopted non-competitive diffusion baselines such as DDPM [24], b) did not consider pre-trained feature extractors for clustering [20], c) did not use the optimal cluster granularity [29]. For instance, Hu et al. [29] used 400 clusters on CIFAR100 using DINO ViT-B for clustering, while Fig. 6.3 shows that 200 clusters lead to a superior FID using C-EDM (11.2% relative improvement).

Comparison with state-of-the-art GT conditional generative models. Intriguingly, using C-EDM with C_V clusters, we report small improvements compared to GT label conditioning on CIFAR10 and CIFAR100 in Table 6.1. For instance, 4.77% mean relative improvement in FID using TEMI. Even though cluster-conditioning is primarily designed for unconditional generation, such as FFHQ, we demonstrate that it can match or outperform GT labels, which showcases the effectiveness of C-EDM.

Sample efficiency. In addition to the reported gains in FID, we study the sample efficiency of C-EDM during training across datasets in Fig. 6.2. On CIFAR10 and CIFAR100, the training sample efficiency for $C = C_V$ compared to the unconditional model ($C = 1$) peaks at 5×, where C-EDM with $M_{img} = 40$ outperforms the unconditional model at $M_{img} = 200$. On FFHQ-64, which is not a classification dataset, we report a sample efficiency of 3.3×. TEMI and k-means demonstrate identical sample efficiency compared to the unconditional model. Intuitively, a more informative conditioning signal enables learning the data distribution faster. Upon visual inspection, we could identify FFHQ clusters for $C = C_V$ with easily distinguishable visual patterns, such as groups of images with beanies, smiling faces, glasses, sunglasses, hats, and kids (see appendix). Finally, we find that the sample efficiency of C-EDM against GT conditional EDM heavily depends on the quality and granularity of the GT labels.

	C		
CIFAR100	20	100	200
Human annotation (GT)	3.10	2.41	-
CLIP pseudo-labels	3.38	2.42	-
k-means clusters	3.09	2.41	2.36
TEMI clusters	**2.93**	**2.37**	**2.31**

Table 6.2: FID for different grouping methods as a condition to EDM ($M_{img} = 100$). We use the 20 GT superclasses and 100 GT labels of CIFAR100. CLIP pseudo-labels are computed using zero-shot classification pseudo-labels based on the GT label names.

D.3 Cluster utilization ratio and discovered upper bounds

We depict how the TEMI cluster utilization r_C changes in tandem with FID for different numbers of clusters in Fig. 6.3. Crucially, this approach also applies to FFHQ, where no GT labels exist. Moreover, we can discern certain patterns across datasets: (i) C_V always has a high utilization ratio, (ii) the majority of experiments with $C < C_{max}$ outperform the unconditional model at $M_{img} = 100$, and (iii) $C_V > C_{GT}$ for CIFAR10 and CIFAR100, which is in line with [29]. Even though the choice of $\alpha = 0.96$ is not guaranteed to be generally applicable and is based on empirical evidence, a more or less strict choice can be used based on the practitioner's computational budget. The introduced upper bound can be computed with negligible computational overhead and without access to GT labels or training the generative model, which on large scales (ImageNet) can require up to 4MWh per experiment [35].

D.4 Investigating the connection between clustering and cluster-conditional image synthesis

FID and ANMI. On image clustering benchmarks, TEMI ($\gamma = 0.6$) outperforms k-means, where we respectively measure an ANMI of 60.6% versus 59.2% on CIFAR10 and 72.3% versus 67.6% on CIFAR100. However, their generative performance displays negligible differences in terms of FID for $C = C_V$. We emphasize that the imbalanced clusters of k-means are penalized when computing clustering metrics such as ANMI on balanced classification datasets, while during image synthesis, this is naturally mitigated by sampling from $q(c)$. Table 6.1 suggests that imbalanced cluster assignments are beneficial for unlabelled datasets such as FFHQ, where we report a relative gain of 4.8% using k-means clusters compared to TEMI.

C-EDM matches EDM when $C = C_{GT}$. In Fig. 6.3, we observe that the GT label conditioning closely follows the FID of TEMI clusters for $C = C_{GT}$. In Table 6.2, we leverage the two annotation levels of CIFAR100 [40], specifically the 20 GT superclasses and the 100 GT labels to benchmark how different grouping methods perform in image synthesis. Apart from image clustering, we create text-based pseudo-labels with CLIP (OpenCLIP ViT-G/14 [31]) similar to [46]. Then, CLIP pseudo-labels are derived from zero-shot classification [54] using the arg max of the text-image similarity after softmax [1].

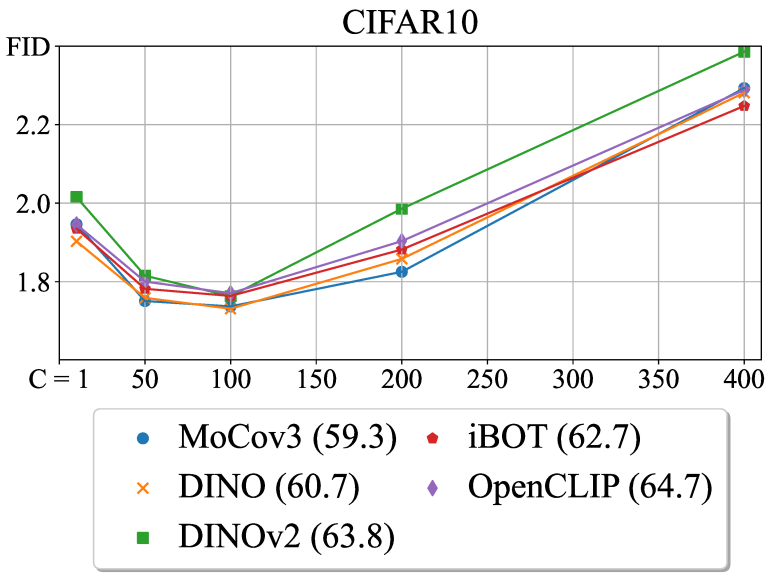
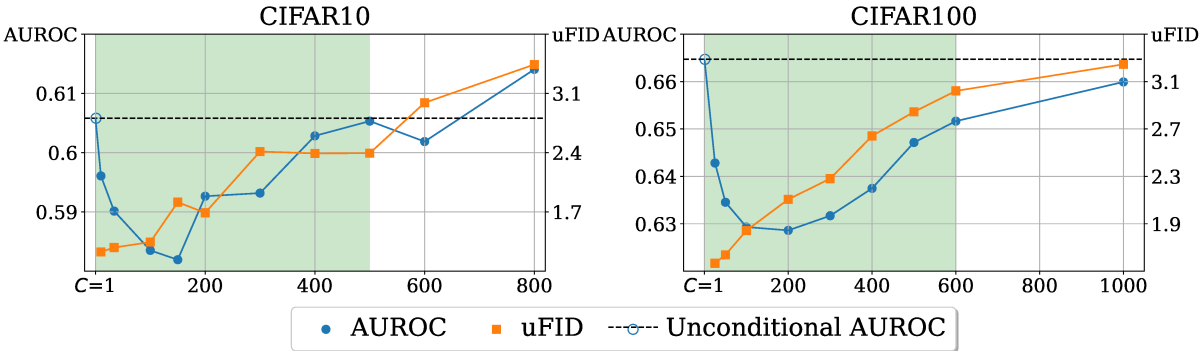


Figure 6.4: FID (*y-axis*) across different numbers of clusters C (*x-axis*) using C-EDM with TEMI with different feature extractors. The ANMI is shown in parentheses for C_V=100.

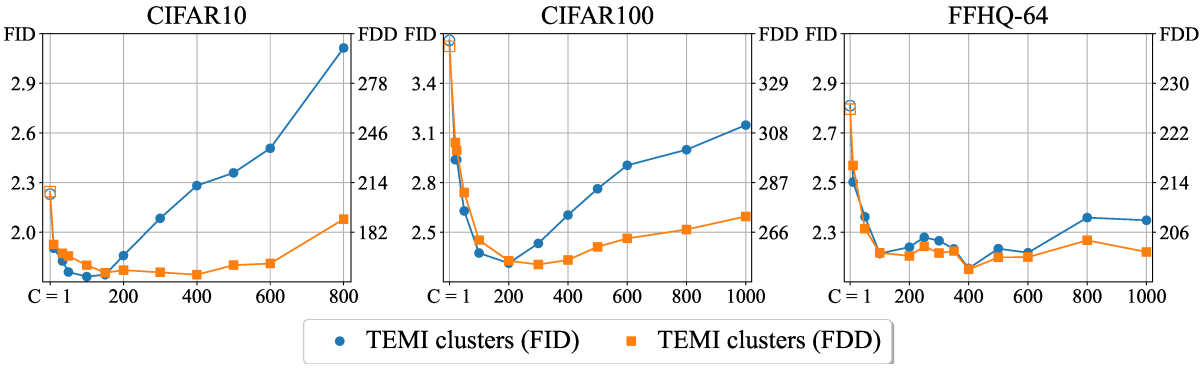
For reference, we provide the FID with the number of visual groups $C_V = 200$ for k-means and TEMI clusters. Interestingly, all methods attain a similar FID for $C_{GT} = 100$, while TEMI achieves the best FID only for $C = 20$. This suggests that generative performance is mostly invariant to the grouping method, including human annotation, given a feature extractor that captures general visual concepts. Next, we investigate the impact of the pre-trained feature extractor.

The number of visual groups is insensitive to the chosen feature extractor. The broader adoption of the discovered C_V for each dataset requires our analysis to be insensitive to the chosen feature extractor. In Fig. 6.4, we show that the pre-trained models considered in this work achieve similar FIDs across different numbers of clusters even though the ANMI varies up to 5.4% on CIFAR10. Feature extractors differ in terms of the number of parameters (86M up to 1.8 billion), pre-training dataset size (1.2M up to 2B), and objective (MoCO[16], DINO[12], iBOT[75], CLIP [31, 54]). Our finding is consistent with concurrent work on the scale of ImageNet [44] that shows that linear probing accuracy is not indicative of generative performance for feature-based conditioning. The fact that differences in discriminative performance do not translate to improvements in FID is in contrast with Hu et al. [29].

Dataset-specific features do not improve generative performance. To realize dataset-dependent adaptation of the feature extractor, we perform self-supervised fine-tuning on CIFAR10 and CIFAR100, starting from the ImageNet weights in Table 6.3. We train all layers for 15 epochs similar to [57] using the DINO framework, resulting in "*DINO adapted weights*". We find large improvements in classification accuracy and ANMI, which is in line with the fact that the learned features are more dataset-specific [55, 57, 71]. However, the gains in classification and clustering do not translate into conditional image generation, which suggests that image synthesis requires features that capture general visual concepts.



(a) For the computation of AUROC, we use the official test splits.



(b) The non-filled markers indicate the unconditional EDM.

Figure 6.5: **a)** Top 1-NN cosine similarity AUROC (*left y-axis*) and Frechet distance between the C-EDM and unconditional samples (uFID) for different cluster sizes C (*x-axis*). **b)** FID (*left y-axis*) and FDD (*right y-axis*) versus number of clusters (*x-axis*) using C-EDM at $M_{img} = 100$.

E Discussion

Highly fine-grained clusters lead to out-of-distribution samples. To quantify the degree of out-of-distribution for the generated C-EDM samples, we measure the AUROC using the top-1 NN cosine similarity (1-NN) using DINO ViT-B [66] and compare with the test split of CIFAR10 and CIFAR100 (Fig. 6.5a). We note that 1-NN is independent of the data distribution of the generated samples and thus does not account for diversity. The highest C considered for CIFAR10 and CIFAR100 produces the highest AUROC, suggesting that the features of the generated samples are pushed away from the training data. We hypothesize that this behavior originates from highly specialized clusters that cannot always be generated from the initial noise, resulting in visual artifacts. In parallel, we measure how similar the C-EDM samples are compared to the unconditional ones using FID. We call this metric *unconditional FID (uFID)*. We observe that uFID increases as C increases, suggesting that the generated samples for larger C are farther away in feature space than the unconditional samples.

The clusters' granularity level determines conditional generative performance. The quality of feature representations is typically determined by linear separability w.r.t. GT labels [50]. However, our experimental analysis shows that in image

C-EDM+TEMI	CIFAR10 ($C = 100$)				CIFAR100 ($C = 200$)			
	ACC↑	ANMI↑	FID↓	FDD↓	ACC↑	ANMI↑	FID↓	FDD↓
ImageNet weights	96.4	60.6	**1.73**	153.5	82.3	72.3	**2.31**	**250.7**
Adapted weights	**98.5**	**63.41**	1.74	**152.8**	**87.8**	**78.2**	2.39	250.8

Table 6.3: Effect of self-supervised fine-tuning (adapted weights) on the pre-trained DINO ViT-B. ACC stands for the 20-NN classification accuracy using the DINO features. We measure FID and FDD for both models at $M_{img} = 100$.

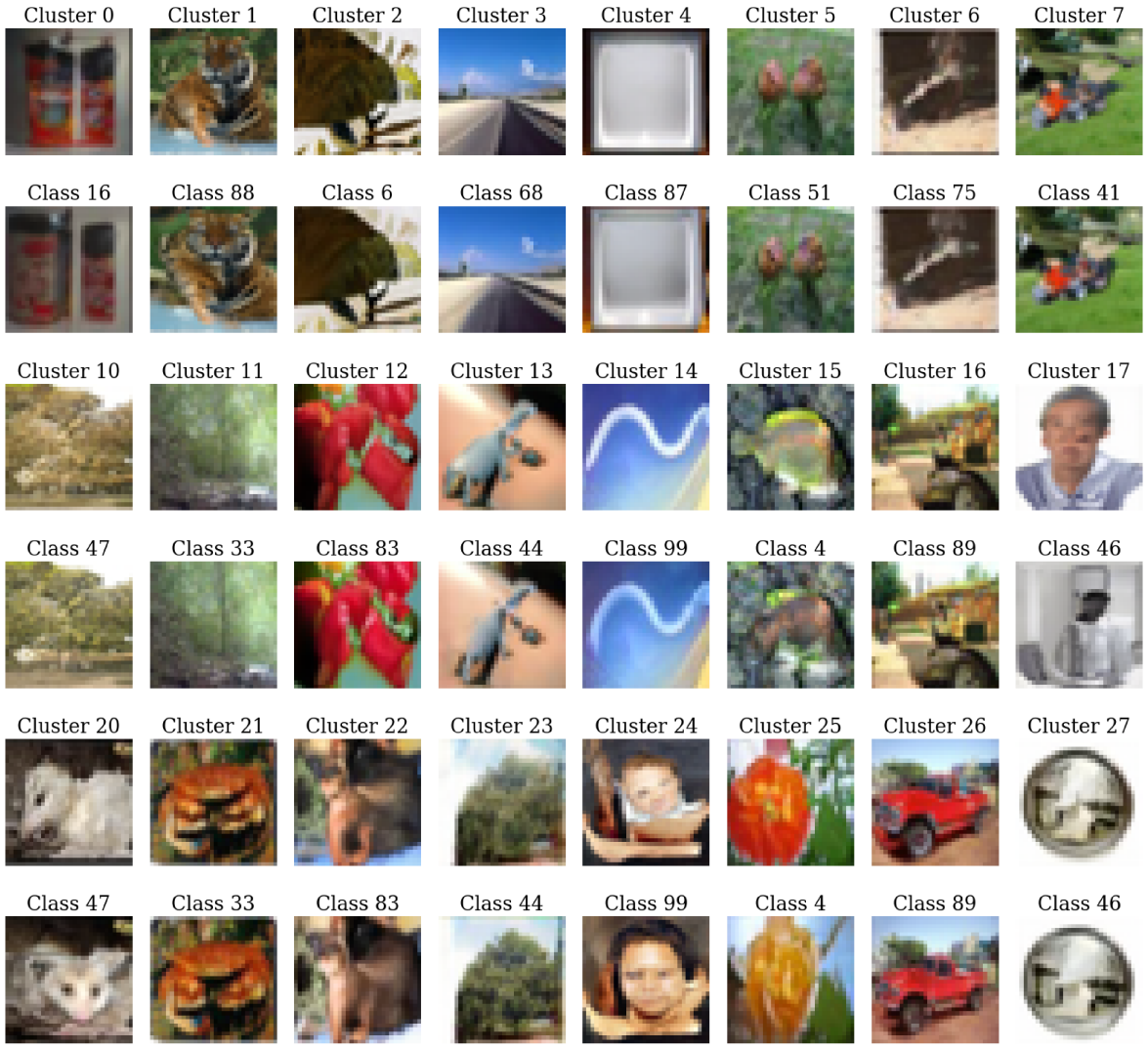


Figure 6.6: Comparing C-EDM samples generated with learned cluster assignments (top row, C=100) to EDM samples generated with GT labels, on CIFAR100. The clusters are mapped to GT classes using the Hungarian algorithm.

Figure 6.7: Visual comparison between C-EDM (C=400) and unconditional EDM (Uncond.) at $M_{img} = 100M$ on FFHQ at 128x128.

synthesis, the quality of image-level conditioning primarily depends on the granularity of the cluster assignments (Table 6.2 and Fig. 6.3). Overall, we find no significant connection between the task of feature-based clustering and cluster-conditional image generation (Fig. 6.4 and Table 6.3). The pre-trained model and the categorization method do not severely affect generative performance for the considered datasets.

Generative metric and optimal number of clusters. Here, we investigate the impact of the choice of generative metric by comparing FDD and FID. We highlight that the only difference is that FID and FDD use InceptionV3 and DINOv2 features, respectively. The cross-dataset evaluation in Fig. 6.5b shows that FID and FDD do not always agree with respect to the number of visual groups. The largest disagreement is observed on CIFAR10. Still, the discovered upper bound C_{max} always includes the number of visual groups. Finally, we measured additional feature-based metrics such as precision, recall, diversity, coverage, MSS, and inception score without success. Similar to [17, 42, 51, 32, 65], we argue that new generative metrics must be developed.

Visual comparison of C-EDM to EDM. In Fig. 6.6, we map clusters to their respective CIFAR100 classes using the Hungarian mapping [41]. We then produce samples with C-EDM and GT conditional EDM using the same noise. Given the same initial noise and the cluster that is mapped to its respective GT class, we observe a lot of visual similarities in the images, even though the two models (C-EDM and EDM) have different

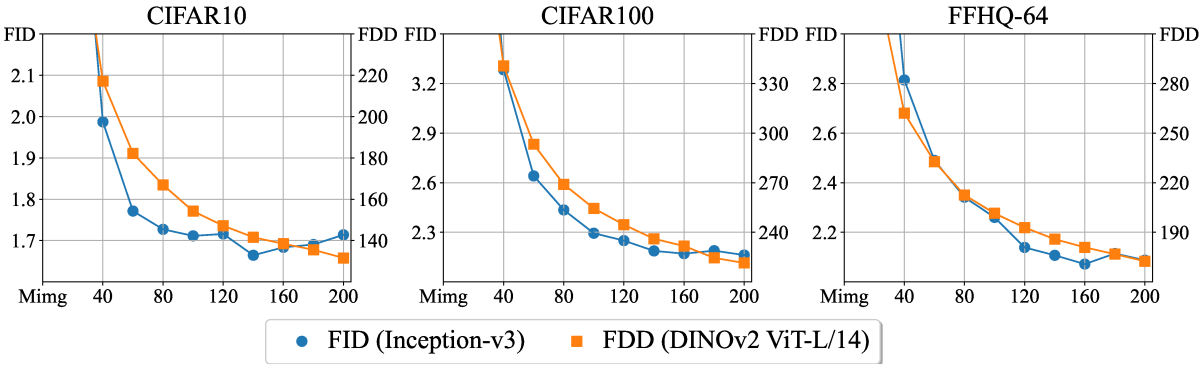


Figure 6.8: FID score (*y-axis, left*) and FDD (*y-axis, right*) during training samples seen (M_{img}, *x-axis*). We used $C_V = 100, 200, 400$ for CIFAR10, CIFAR100 and FFHQ-64 respectively.

weights and have been trained with different types of conditioning. This observation corroborates with Table 6.2 and Fig. 6.2.

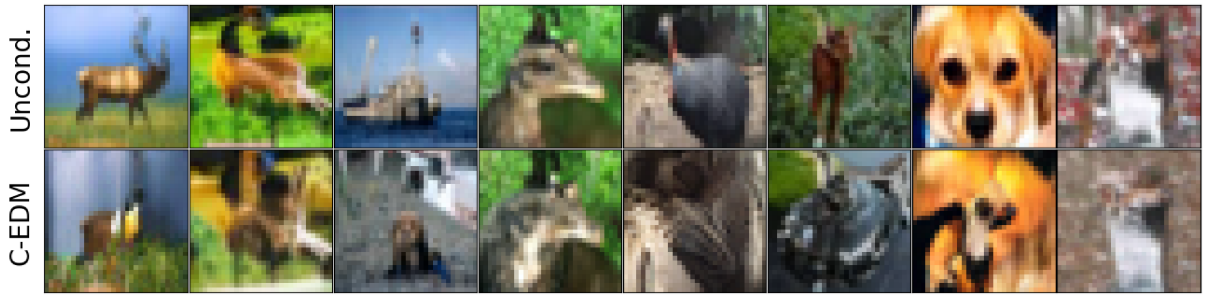
In Fig. 6.7, we visualize cluster-conditioned and unconditional FFHQ samples at $M_{img} = 100M$. In Fig. 6.10, we visualize C-EDM samples generated from the same initial noise on FFHQ-64 for diffusion models trained with varying cluster granularities. Each noise gets a condition sampled from $p(c)$. Similar to our quantitative analysis, the generated images from small cluster sizes are closer to the unconditional prediction.

FID and FDD across training iterations In Fig. 6.8, we report FID and FDD across training using C-EDM with TEMI clusters. We notice that FID tends to saturate faster than FDD and fluctuates more between checkpoints. FDD keeps decreasing monotonically, with minimal fluctuation and always prefers the samples at $M_{img} = 200$. Since both metrics compute the Frechet distance, these tendencies can only be attributed to the supervised InceptionV3 features. Even though the study of generative metrics is out of the scope of this work and a human evaluation is necessary as in [65], we hope that our findings w.r.t. cluster-conditioning can facilitate future works.

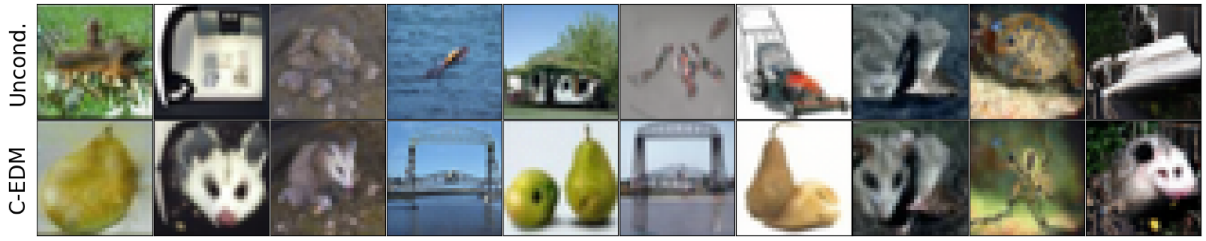
Measuring the confidence of the generated samples using TEMI. In Fig. 6.9, we show the generated examples with the lowest and highest maximum softmax probability [22] of the TEMI head as a measure of confidence. For comparison, we show unconditional samples generated using the same initial noise in the denoising process. Visual inspection shows that low-confidence C-EDM samples do not have coherent semantics compared to the unconditional ones, leading to inferior image quality. We hypothesize that the sampled condition for the low-confidence samples is in conflict with the existing patterns in the initial noise. Increasing the number of conditions likely leads to worse image-condition alignment [60, 25, 27]. We argue that internal guidance methods could be employed to increase the image quality in such cases, which is left for future work.

By contrast, highly confident C-EDM samples show more clearly defined semantics than unconditional ones. When the low frequencies, such as the object's shape, remain intact, cluster conditioning aids in refining local pixel patterns. Confident C-EDM samples consist of simple pixel patterns that are easy to generate, such as a white background. Confidence can also be leveraged in future works in rejection sampling schemes [14]. More samples are provided in the appendix.

(a) High-confidence C-EDM *CIFAR10* samples



(b) Low-confidence C-EDM *CIFAR10* samples



(c) High-confidence C-EDM *CIFAR100* samples

(d) Low-confidence C-EDM *CIFAR100* samples

Figure 6.9: Generated high (a,c) and low (b,d) confidence samples using the trained C-EDM. The top row depicts the unconditional (*Uncond.*) samples, while the bottom row shows the generated samples using *C-EDM* with TEMI ($C = 100$). Images on the same column are produced with the same initial noise.

Image synthesis beyond ImageNet. ImageNet is currently the largest labeled public dataset, and a single experiment using a recent state-of-the-art diffusion model on ImageNet requires up to 4MWh at 512^2 resolution [35]. Based on our experiments, clusters match or outperform the human-derived labels on image generation by estimating the visual groups. Using the introduced upper bound, the search space of the visual groups is significantly reduced with minimal computational overhead, while no further hyperparameter tuning is required. Therefore, it allows future works to incorporate unlabelled data and experiment at scales beyond ImageNet while being sample efficient. Additionally, the sample efficiency compared to noisy or non-mutually exclusive labels could be investigated in future work.

F Conclusion

In this paper, a systematic empirical study was conducted focusing on conditioning diffusion models with cluster assignments. It was demonstrated that cluster conditioning achieves state-of-the-art FID on three generative benchmarks while attaining strong sample efficiency. To reduce the search space for estimating the visual groups of the dataset, a novel method that computes an upper cluster bound based solely on clustering was proposed. Finally, our experimental study indicates that generative performance using cluster assignments depends primarily on the granularity of the assignments.

References

[1] Nikolas Adaloglou, Felix Michels, Tim Kaiser, and Markus Kollmann. Adapting contrastive language-image pretrained (clip) models for out-of-distribution detection. *arXiv e-prints*, pages arXiv–2303, 2023.

[2] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[3] Nikolas Adaloglou, Felix Michels, Kaspar Senft, Diana Petrusheva, and Markus Kollmann. Scaling up deep clustering methods beyond imagenet-1k. *arXiv preprint arXiv:2406.01203*, 2024.

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[5] Björn Barz and Joachim Denzler. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, 2020.

[6] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[7] James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

[8] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[9] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about, 2022.

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[13] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

[14] George Casella, Christian P Robert, and Martin T Wells. Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, pages 342–347, 2004.

[15] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[17] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[20] Bao et al. Why are conditional generative models better than unconditional ones? In *NeurIPS Workshop on Score-Based Methods*, 2022.

[21] Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. On partial prototype collapse in clustering-based self-supervised learning, 2024.

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[26] Laura Hollink, Aysenur Bilgin, and Jacco Van Ossenbruggen. Is it a fruit, an apple or a granny smith? predicting the basic level in a concept hierarchy. *arXiv preprint arXiv:1910.12619*, 2019.

[27] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7462–7471, October 2023.

[28] Vincent Tao Hu, Yunlu Chen, Mathilde Caron, Yuki M Asano, Cees GM Snoek, and Bjorn Ommer. Guided diffusion from self-supervised diffusion features. *arXiv preprint arXiv:2312.08825*, 2023.

[29] Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. Self-guided diffusion models. In *CVPR*, 2023.

[30] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[31] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[32] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023.

[33] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[34] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[35] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2023.

[36] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.

[37] David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.

[38] Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.

[39] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[40] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[41] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[42] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

[43] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[44] Tianhong Li, Dina Katabi, and Kaiming He. Self-conditioned image generation via generating representations. *arXiv preprint*, 2023.

[45] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14286–14295, 2020.

[46] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022.

[47] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[48] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[50] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2023.

[51] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023.

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[55] Nima Rafiee, Rahil Gholamipoor, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, and Markus Kollmann. Self-supervised anomaly detection by self-distillation and negative sampling. In *International Conference on Artificial Neural Networks*, pages 459–470. Springer, 2022.

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[57] Tal Reiss, Niv Cohen, Eliahu Horwitz, Ron Abutbul, and Yedid Hoshen. Anomaly detection requires better representations. In *European Conference on Computer Vision*, pages 56–68. Springer, 2022.

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[59] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradely, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023.

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[61] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval, 2022.

[62] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

[63] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[64] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[65] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel

Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[66] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[67] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[68] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[69] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[70] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[71] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

[72] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[73] Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. Pfgm++: Unlocking the potential of physics-inspired generative models. *arXiv preprint arXiv:2302.04265*, 2023.

[74] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[75] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[76] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022.

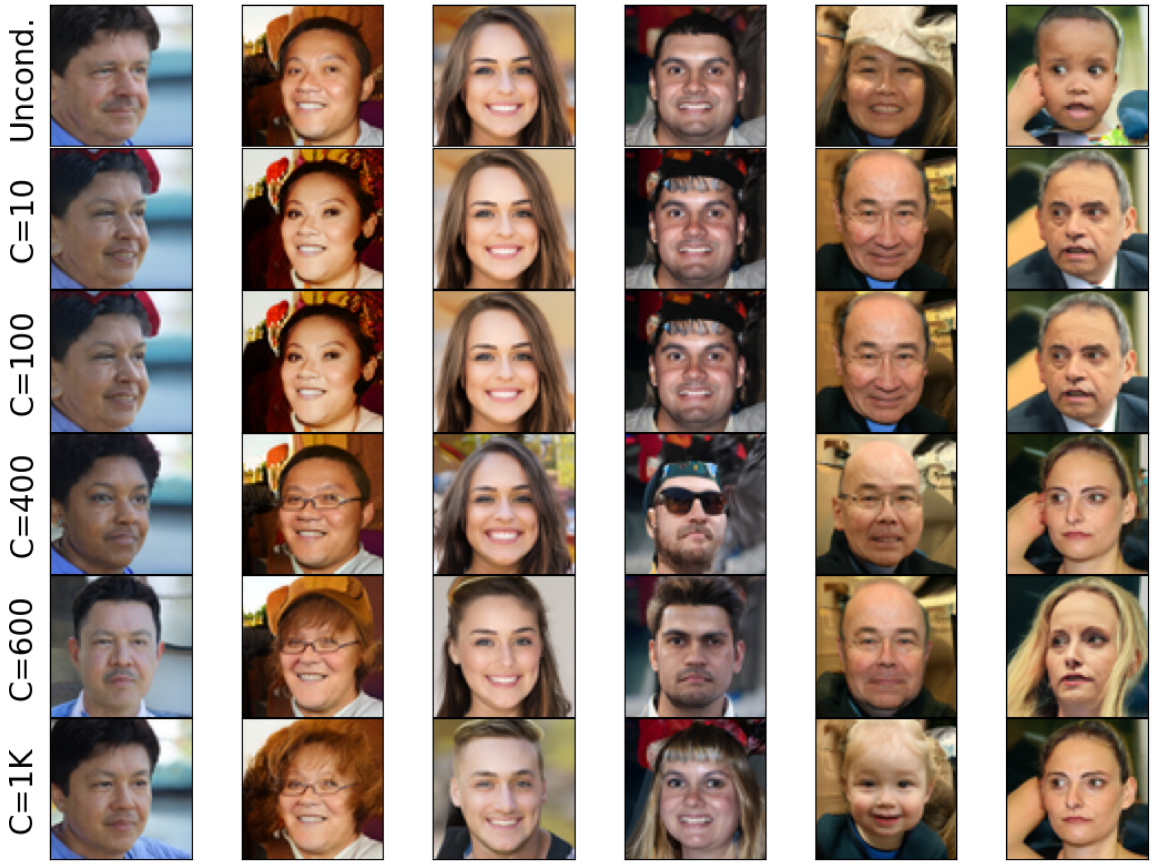


Figure 6.10: Visualizing *generated* images from FFHQ-64 using CEDM for different number of clusters C with the same random noise. We use deterministic noise sampling. Each noise gets a condition sampled from $p(c)$ for each individual clusters.

Appendix

App. I Visualizations

In Fig. 6.11, we visualize *real* training FFHQ images that are grouped in the same TEMI cluster using the DINO features. We visually identify groups with shared characteristics such as sunglasses, hats, beanies, pictures of infants, and pictures of speakers talking to a microphone.

App. II What about a lower cluster bound?

Starting with a high overestimation of the number of clusters (e.g. $1K$ for CIFAR10), we find that TEMI clustering with $\gamma = 1$ utilizes a subset of clusters, which could be used as a lower cluster bound. More precisely, we find a maximum standard deviation of 6.4 for C^u across datasets and feature extractors. Intuitively, C^u is the minimum amount of clusters TEMI (with $\gamma = 1$) uses to group all image pairs. This behavior is analogous to cluster-based self-supervised learning (using image augmentations) [75, 5] and has been recently coined as partial prototype collapse [21]. Nonetheless, the lower bound is more

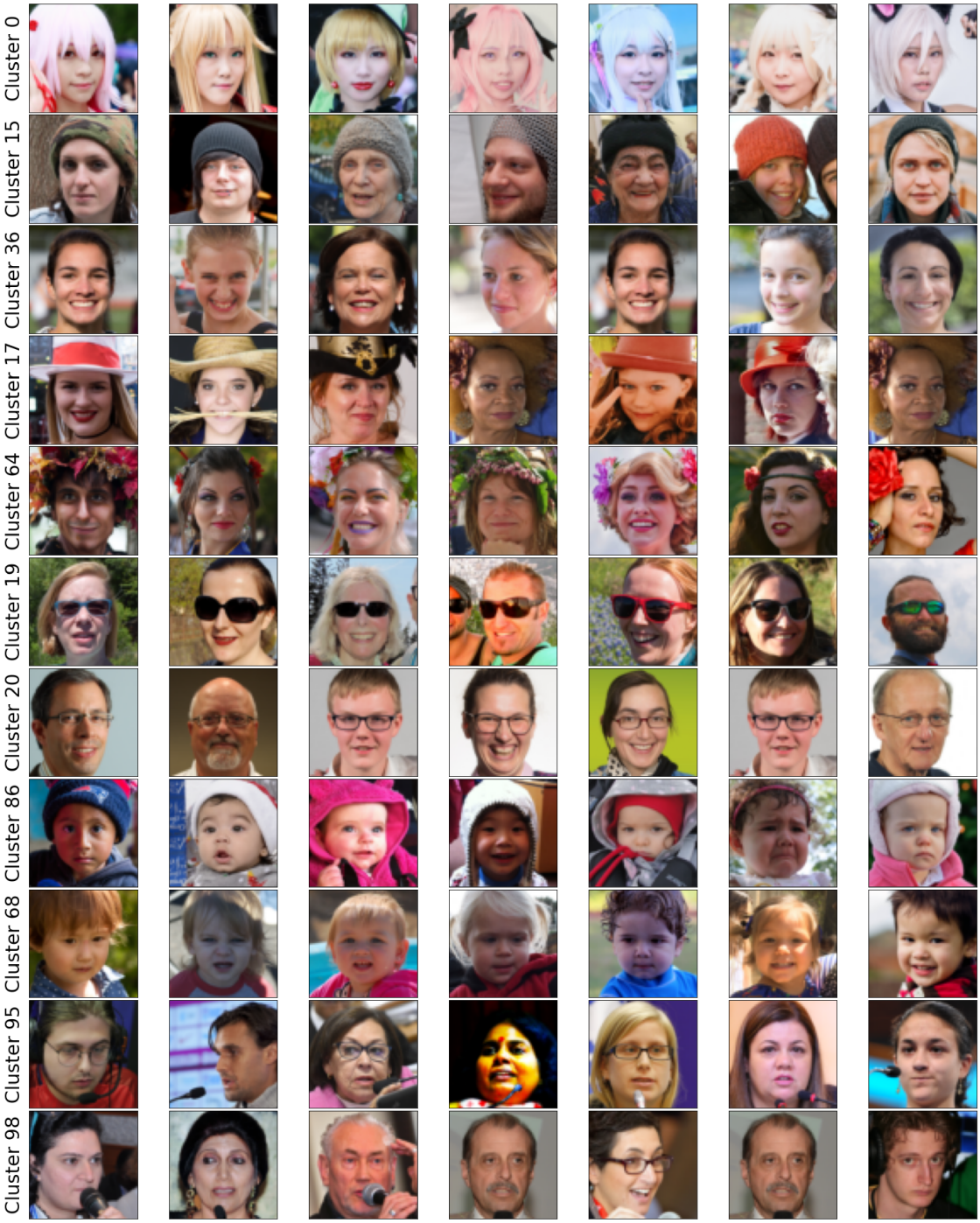


Figure 6.11: Visualizing *training* images from FFHQ that belong to the same TEMI cluster. Images that are grouped into the same cluster are shown in the same row. We use the trained TEMI model with $C_V = 400$ using the DINO backbone. Cluster assignments are picked to illustrate that images with similar visual characteristics are grouped together (i.e., beanies, smiling faces, glasses, hats, kids, etc.). Images are **randomly** sampled from each cluster.

TEMI $\gamma = 1$	CIFAR10 C^u	FFHQ C^u	CIFAR100 C^u
100	33	36	48
400	38	48	48
500	34	54	51
800	40	45	47
1K	34	49	47
2K	48	52	54
5K	28	42	51
Mean	36.4	46.6	49.4
Std	**6.37**	6.16	2.63

TEMI $\gamma = 1$	CIFAR10 C^u
DINO-B	34
MoCOv3-B	39
iBOT-L	45
OpenCLIP-G	47
DINOv2-g	50
Mean	43
Std	6.44

Table 6.4: Number of utilised clusters C^u for different number of input clusters C (*left*) and different backbones (*right*) using TEMI with $\gamma = 1$ and C=500 with the DINO ViT-B/16 backbone. We show the relatively small sensitivity of C^u to the choice of C and backbone; we report a standard deviation of a maximum value of 6.37 across different cluster sizes and 6.44 across backbones on CIFAR10.

applicable to large scales as the measured standard deviation might exclude the optimal granularity for small, highly curated datasets. Due to the above limitation, we leave this for future work.

As depicted in Table 6.4, the utilized number of clusters C^u is not sensitive to the pre-determined number of clusters nor the choice of backbone for TEMI clustering when $\gamma = 1$.

App. III Additional implementation details and hyperparameters

When searching for C_V, we evaluate EDM after training with $M_{img} = 100$ and for $M_{img} = 200$ once C_V is found. We only report k-means cluster conditioning with $k = C_V$. All our reported FID and FDD values are averages over three runs of 50k images each, each with different random seeds. Below, we show the hyperparameters we used for all datasets to enable reproducibility. We always use the average FID and FDD for three sets of 50K generated images. The used hyperparameters can be found in Table 6.5 .

To assign the CLIP pseudo-labels to the training set, we compute the cosine similarity of the image and label embeddings using openclip ViT-G/14 [31]. The label embeddings use prompt ensembling and use the five prompts: a photo of a <label>, a blurry photo of a <label>, a photo of many <label>, a photo of the large <label>, and a photo of the small <label>.

Table 6.5: Hyperparameters used for training EDM and C-EDM. **Bold** signifies that the value is changing across datasets. All other parameters of the training setup were identical to the specifications of Karras et. al [34], which are detailed there.

Hyperparameter Hyperparameter	CIFAR10/ CIFAR100	FFHQ-64/ AFHQ-64
Optimization		
optimizer	Adam	Adam
learning rate	0.001	0.001
betas	0.9, 0.999	0.9, 0.999
batch size	1024	512
FP16	true	true
SongUNet		
model channels	128	128
channel multiplier	2-2-2	1-2-2-2
dropout	13%	5% / 25%
Augmentation		
augment dim	9	9
probability	12%	15%

7. Scaling Up Feature-Based Image Clustering Methods Beyond ImageNet-1K

Outline

A Introduction 1
B Related Work 2
C Background, Materials and Methods 3
D New Clustering Benchmarks 5
E Experimental Results 7
F Discussion and Limitations 10
G Conclusion 12
References 13
Appendix 18

Bibliographic Information

Adaloglou, N., Michels, F., Senft*, K., Petrusheva*, D., & Kollmann, M. (2024). Scaling Up Deep Clustering Methods Beyond ImageNet-1K. arXiv preprint arXiv:2406.01203.

Corresponding author: *adaloglo@hhu.de*

Copyright Notice

CC BY 4.0 Attribution 4.0 International. The latest preprint of this article is available at `https://arxiv.org/abs/2406.01203`. Clarification of the copyright was adjusted according to the guidelines of the publisher.

Acknowledgements

The author would like to thank i) the co-author Kaspar Senft for his constant feedback and help with revising the manuscript, ii) the co-author Felix Michels for adjusting

and speeding up the code for large-scale benchmarks on ImageNet21K, and iii) Diana Petrusheva for prototyping the early version of the zero-shot label refining method.

Individual Contributions

The author came up with the idea of scaling up and benchmarking feature-based clustering algorithms. To this end, the ImageNet-21K dataset was selected and preprocessed by the author. Furthermore, the author proposed evaluating the performance of clustering algorithms with respect to dataset-specific factors, such as class imbalance, and was primarily responsible for developing the corresponding benchmarks. In collaboration with Diana Petrusheva and Kaspar Senft, the author proposed a hierarchical zero-shot label refinement method. The latter method computes more fine-grained labels by leveraging a semantic hierarchy (i.e. WordNet). The author co-developed the multi-label benchmarks along with Kaspar Senft. Finally, the author conducted a significant portion of the experiments and wrote the manuscript.

Abstract

Image clustering methods are often evaluated on balanced classification datasets up to the scale of ImageNet-1K. This limits fair comparisons to standard baselines like k-means clustering, which are frequently used with larger, uncurated datasets. In this work, we explore the performance of feature-based clustering approaches on large-scale benchmarks while disentangling the impact of the following data-related factors: i) class imbalance, ii) class granularity, iii) the presence of easily recognizable classes, and iv) the ability to capture multiple classes. To facilitate this analysis, we introduce several new benchmarks derived from ImageNet21K. We find that recent deep learning feature-based methods generally outperform k-means across most of our benchmarks. However, the performance gap diminishes as the number of classes increases, regardless of class granularity or non-mutual exclusive classes. Interestingly, k-means underperforms on easy-to-classify benchmarks by large margins. Finally, we demonstrate that combining a recent feature extractor with a well-tuned feature-based clustering method achieves a state-of-the-art clustering accuracy of 69.8% on ImageNet-1K without external data or labels.

A Introduction

For over a decade, the ImageNet-1K benchmark [18, 45] has served as the core downstream dataset in a plethora of computer vision tasks. Examples include image classification [25], representation learning [15], semi-supervised learning [49], and recently image clustering [62]. Image clustering, also known as unsupervised image classification, refers to algorithmically grouping visual stimuli into discrete concepts called clusters [10, 2]. Grouping images without human supervision has many applications, including unsupervised out-of-distribution detection [47], image generation [8], and large-scale dataset pruning [40, 1].

Motivation. Advancements in representation learning [7, 12, 41] have led to the emergence of feature-based clustering. To date, the most widely established clustering approach in feature space is k-means [35] due to its scalable nature. Oquab et al. [40] scaled up feature-based k-means [35] to 100K clusters using a billion-scale proprietary dataset, which is not publicly available for benchmarking. Concurrently, feature-based clustering methods such as TEMI [3] have achieved state-of-the-art clustering performance on ImageNet-1K (1.2M samples) [18, 45], outperforming k-means by significant margins. This discrepancy in dataset scales raises questions about the applicability of recent methods beyond the scope of balanced classification datasets (ImageNet-1K), as well as their comparability to k-means in real-world, large-scale scenarios. To address this, we start by first presenting the key inherent and non-inherent limitations regarding image clustering.

Inherent limitations. The first inherent limitation of parametric clustering is that the valid computation of the clustering accuracy requires setting the number of clusters to the number of human-derived classes, the *ground truth (GT)*. Second, similar to image classification, real-world images frequently contain multiple objects of interest, which is not captured when considering only the top 1 class prediction. Third, biases in the data annotation procedure can lead to systematic inaccuracies [11], even at small data regimes [9]. Human annotators may assign a particular label that seems plausible when presented in isolation or create visually indistinguishable groups based on arbitrary class distinctions (e.g. the class "laptop computer" and "notebook computer" coexist in ImageNet-1K) [11, 10]. Finally, unlike supervised approaches, clustering algorithms cannot accurately capture highly coarse labels solely based on visual input, such as carnivoran mammals [2, 44].

Non-inherent limitations. More importantly, current benchmarks have specific constraints that are non-inherent limitations of image clustering. First, the majority of methods are being developed on small-scale datasets such as CIFAR10 and CIFAR100 [33], with only a limited number of methods reporting results on ImageNet-1K [52, 6, 3]. It is thus unclear whether these methods apply to large-scale benchmarks, often considering their non-scalable nature. Spectral clustering with an $O(N^3)$ time complexity [38] is one example, where N is the number of samples. Second, existing benchmarks are primarily balanced classification datasets, wherein all samples are uniformly distributed among the GT labels. While deep learning-based clustering methods leverage this information using regularization techniques, classical machine learning methods (e.g. k-means) do not have this flexibility and tend to create imbalanced clusters [52]. New challenging benchmarks that mimic real-world scenarios would aid in bridging the gap between ImageNet-1K and

billion-scale proprietary datasets.

In this paper, we perform a comprehensive experimental study on large-scale clustering methods and benchmarks. We create new clustering benchmarks based on ImageNet21K, focusing on the effect of individual factors: i) class imbalance, ii) class granularity, iii) easily "classifiable" classes, and iv) the ability to capture multiple labels. We show that evaluating only balanced classification datasets treats k-means unfairly, yet deep learning-based clustering methods still outperform k-means across most benchmarks besides ImageNet-1K. The gap, however, diminishes on large-scale benchmarks with more than 7.3K classes. We find that i) k-means attains inferior performance on easy-to-classify benchmarks by large margins, and ii) a recent feature extractor [4] with a well-tuned deep learning-based clustering method [3] achieves a state-of-the-art clustering accuracy of 69.8% on ImageNet-1K without external data or labels, surpassing the accuracy of a supervised GoogLeNet [51].

B Related Work

B.1 Deep learning-based image clustering

Deep learning-based image clustering consists of learning the label-related representations and the cluster assignments using neural networks [13, 3], either simultaneously (single-stage methods) [62, 6] or sequentially (multi-stage methods) [52].

Single-stage methods. Single-stage methods [57, 63, 62, 6] are computationally demanding and have been overshadowed by advancements in visual representation learning [7, 12]. For instance, in SELA [62], the authors design a two-step framework, which alternates between estimating the pseudo-label assignment matrix and representation learning. In [6], the authors present a single-stage end-to-end method that employs a variant of the cross-entropy loss. However, such single-stage methods require dataset-dependent hyperparameter tuning and are computationally expensive to iterate on large scales.

Multi-stage methods. Recent state-of-the-art methods [52, 3] primarily operate on image representations rather than focusing on learning specific representations for clustering. We refer to this category as *feature-based image clustering approaches*. The simplest approach is feature-based k-means, which usually creates imbalanced clusters [52].

Contrastive learning [15] has been a major breakthrough for image clustering. SCAN [52] was the first large-scale method to isolate representation learning from learning the cluster assignments. By first employing contrastive learning on the training data, Van Gansbeke et al. [52] show that nearest neighbors (NNs) in feature space likely share the same label. Based on the same principle, Adaloglou et al. [3] have recently shown that a) features from multiple pre-trained feature extractors (i.e. DINO, MSN [7]) can be used and b) learning the cluster assignments does not require image augmentations as opposed to SCAN. More recent approaches focus on learning task-specific representations by sequentially combining masked image modeling with contrastive learning [4, 48], which has been recently coined as pre-pretraining [48].

As we focus on feature-based methods that are scalable to large-scale datasets, we investigate how to learn the cluster assignments. Since we adopt pre-trained feature

extractors, our work is closer to TEMI [3], an alternative to feature-based k-means. Nevertheless, deep learning-based clustering methods have only been applied up to the scale of ImageNet-1K ($\approx 1.2M$ images). Using recent advancements in similarity search and a distributed implementation of k-means [32, 20, 40], feature-based k-means has been scaled to billion scale proprietary datasets. However, this dataset is not publicly available for benchmarking. The divergence in the above data regimes makes it unclear whether other clustering methods can be scaled beyond ImageNet-1K, which we aim to address in this work using ImageNet21K.

B.2 Leveraging ImageNet21K class structure

The set of ImageNet21K classes and its semantic tree has been used in a plethora of downstream applications, such as hierarchical image classification [39], object detection [65] and zero-shot learning [56, 14, 61]. In [65], the authors apply CLIP on the predicted bounding boxes to generate diverse labels based on ImageNet21K for object detection. To increase the generalization capability to unseen classes (test set), Yi et al. [61] exploit the semantic structure of ImageNet21K during vision-language training for zero-shot learning. In [39], the authors demonstrate the superior performance of CLIP (zero-shot classification) leveraging more fine-grained classes than the GT. The discussed works have explored the ImageNet21K semantic structure to boost downstream task performance. Instead, we leverage the ImageNet21K semantic tree to vary the GT class granularity.

C Background, Materials and Methods

Existing large-scale benchmarks and pre-trained models ImageNet-1K [18, 45] consists of C=1000 balanced classes ($p(c) = \mathcal{U}(\{1, .., C\})$) and is the largest clustering benchmark to date. In the context of image clustering, the only existing artificial class imbalance is halving the number of samples from the odd class indices [19], which can only be applied to balanced data. For completeness, we apply this imbalance to ImageNet-1K instead of CIFAR [33]. Meanwhile, we do not modify the validation set to quantify the performance degradation attributed to the imbalance. We refer to this benchmark as *ODD*. Finally, we use the reassessed multi-label validation set of ImageNet-1K called "ReaL" [11] that consists of 50K samples, namely ImageNet-1K ReaL.

The new benchmarks are based on the preprocessed version of ImageNet21K (winter version 2021) as in [42] with $\approx$ 11.06M samples that are non-uniformly distributed across 10450 non-mutually exclusive classes. These classes can be organized in the WordNet semantic tree [37, 59]. To imitate a real-life scenario where GT labels and additional data are scarce, we assume no access to models trained on external data. Unless otherwise specified, we use the iBOT ViT-L [64] pre-trained on ImageNet21K and MAE Refined ViT-H [4] (*MAE-R*) on ImageNet-1K.

C.1 Feature-based image clustering methods

Classical clustering methods do not scale. Given N samples, classical clustering methods become impractical for large-scale datasets like ImageNet[1] where $N > 10^6$.

[1] `https://scikit-learn.org/stable/modules/clustering.html`

The main reason is their reliance on global optimization, which requires processing all available samples N at each step. For instance, spectral clustering has a complexity of $O(N^3)$, agglomerative methods such as WARD [31] run in $O(N^2)$, and DBSCAN also operates in $O(N^2)$. Moreover, these methods typically fit the test data, whereas our focus is on clustering approaches that generalize to unseen samples. To this end, we prioritize scalable alternatives, such as distributed k-means [32, 20] as a baseline, as it efficiently fits the training data while enabling predictions on new data instances.

Deep learning-based methods. Clustering methods that do not leverage pretrained models are generally less competitive [57, 24, 63, 58, 43], challenging to tune, and computationally expensive. Therefore, in addition to k-means, we focus on recent state-of-the-art methods [3], specifically TEMI and SCAN, which are described below. A key advantage of these methods is their optimization using mini-batches, allowing them to scale independently of N. To the best of our knowledge, no other scalable, open-source, feature-based clustering method currently supports datasets exceeding 10^7 samples such as ImageNet21K.

TEMI first mines the nearest neighbors (NN) in feature space for each sample x in the dataset D. During training, TEMI randomly samples x from D and x' from the set of NN of x. TEMI employs a teacher and student head $h_t(\cdot)$ and $h_s(\cdot)$ that have an identical architecture but differ w.r.t. their parameters as the teacher is only updated using an exponential moving average of $h_s(\cdot)$. Given a feature extractor $g(\cdot)$, the student and teacher consist of multiple independent heads $h_s^i(z)$, $h_t^i(z')$, where $i \in \{1, \dots, H\}$ is the head index and $z = g(x)$, $z' = g(x')$. Class probabilities are computed using a softmax function $q_s(c|x)$ and $q_t(c|x')$ and the following loss is minimized

$$\mathcal{L}^i_{\text{TEMI}}(x, x') \coloneqq \underbrace{-\frac{1}{H}\sum_{j=1}^{H}\sum_{c'=1}^{C} q_t^j(c'|x) q_t^j(c'|x')}_{\text{instance weighting}} \log\left(\sum_{c=1}^{C} \frac{\left(q_s^i(c|x) q_t^i(c|x')\right)^\beta}{\tilde{q}_t^i(c)}\right). \tag{7.1}$$

In Eq. (7.1), $\tilde{q_t}^i(c)$ is an estimate of the teacher cluster distribution $\mathbb{E}_{x\sim p_{\text{data}}}[q_t^i(c|x)]$, which can be computed by an exponential moving average over mini-batches B defined as $\tilde{q_t}^i(c) \leftarrow \lambda\, \tilde{q_t}(c) + (1-\lambda)\frac{1}{|B|}\sum_{x\in B} q_t^i(c|x)$ for $\lambda \in (0, 1)$. The scalar $\beta \in (0.5, 1]$ avoids assigning all sample pairs to a single cluster (mode collapse) and is typically set to 0.6.

SCANv2 refers to using the same self-distillation framework as TEMI, but with the semantic clustering loss of SCAN [52],

$$\mathcal{L}^i_{\text{SCAN}}(x, x') \coloneqq -\log \sum_{c=1}^{C} q_s^i(c|x) q_t^i(c|x') \underbrace{-\alpha \sum_{c=1}^{C} \hat{q_s}^i(c) \log \hat{q_s}^i(c)}_{\text{Negative entropy of } \hat{q_s}^i(c)}, \tag{7.2}$$

where $\alpha > 0$ is a scalar and $\hat{q_s}^i(c) = \frac{1}{|B|}\sum_{x\in B} q_s^i(c|x)$ is the current mini-batch estimate

of the student cluster distribution. Unlike the TEMI objective, $\mathcal{L}^i_{\text{SCAN}}(x, x')$ only depends on the head number i.

The entropy term of Eq. (7.2) is upweighted by $\alpha = 5$ and acts conceptually similar to $\tilde{q}_t{}^i(c)$ in Eq. (7.1). Both Eqs. (7.1) and (7.2) are symmetrized and averaged across heads. After training, the head with the lowest training loss is kept for evaluation.

D New Clustering Benchmarks

D.1 Sensitivity to class imbalance

To explore the effects of class imbalance, we first sort the classes according to the number of samples per class (frequency). We then take percentiles s around the median $[50 - s, 50 + s]$ to create frequency-based subsets, details of which can be found in the appendix. As s increases, the number of samples N and classes C also increase. To create an imbalanced subset, we take the middle 10% of classes and add the classes whose number of samples fall in the top 10%. This yields a subset of 2113 classes, which we refer to as *Imbalanced-2K* (Table 7.1). For comparison, we create a similar-size near-balanced subset with $s \approx 10$ yielding the same number of classes centered around the median (*Centered-2k*).

D.2 Sensitivity to the class granularity

Coarse benchmarks. Class labels can be organized in a hierarchical semantic tree [50], such as WordNet [37]. Based on WordNet, 11 hierarchy levels have been identified on ImageNet21K [42]. A hierarchy of 1 corresponds to the most coarse level and 11 to the most fine-grained. Using a semantic tree, one can seamlessly find the linguistic hypernyms (semantic ancestors) for each GT class, such as going from the GT class "chair" to "furniture". By leveraging the ImageNet21K tree, we recursively map the GT classes to their semantic ancestor such that the maximum hierarchy depth is restricted to $d \in [1, .., 9]$, where the majority of the samples lie. Note that we do not exclude any samples at this step and leave classes with a lower hierarchy than d intact. During the learning process of the clustering algorithm, we create as many groups as the number of coarsened labels and evaluate on the validation set with the coarse labels of maximum depth d.

Fine-grained benchmarks. Refining ground truth (GT) labels to more fine-grained classes is more challenging than coarsening them. The simplest way to achieve maximum granularity is to consider only images from classes without linguistic hyponyms (semantic descendants) in the semantic tree, known as leaf classes [60]. Based on their respective trees, we refer to these subsets as *WordNet leaf* and *ImageNet21K leaf*. Unlike ImageNet21K, these subsets are mutually exclusive. To avoid excluding samples from non-leaf classes, we first use CLIP in tandem with the tree to reannotate non-leaf samples. To this end, we perform zero-shot classification on the hyponyms to obtain a new refined label, such as furniture→{chair, desk, shelf, couch}. This procedure is recursively applied, always selecting labels with the highest image-text similarity. We stop once a leaf class is assigned. We refer to this method as leaf hierarchical zero-shot label refining (*leaf HZR*).

Figure 7.1: Reassessed ImageNet21K samples using parent hierarchical zero-shot label refining (p-HZR). The original GT label is shown on top, while each column shows the reassessed label using openCLIP ViT-G [30] in conjunction with the semantic tree. The linguistic hyponyms of the label athletic game are ice hockey, tetherball, water polo, outdoor game, and court game.

However, this approach does not guarantee that an image can always be meaningfully relabeled as one of its hyponyms. Analogously to a human annotator that can decide to keep the current label [23, 22], we add the parent class during zero-shot classification. In the previous example, we perform zero-shot classification using the following class set {chair, desk, shelf, couch, furniture}. By including the parent class, we allow CLIP to assign the same label to the image, partially mitigating the aforementioned issue. As illustrated in Fig. 7.1, not all basketball images are refined to professional basketball. We refer to this setup as parent HZR *(p-HZR)*.

D.3 Easy-to-classify benchmarks

Here, we assess whether easy-to-classify samples are easy to cluster, as measured by high linear class separability, i.e. as measured by probing ACC [53]. The intuition behind this is that if an image is not classified correctly, it is unlikely to be clustered successfully. We choose classes that are recognized with high ACC from models with additional information during pre-training. To realize this, we use a subset of the top-1K class accuracies from three models: 1) a vision-language model (OpenCLIP ViT-G [30]), 2) a visual self-supervised model (DINOv2 ViT-g [40]), and 3) an ImageNet21K-supervised model (ViT-L [55]). More precisely, we measure the top-1K accuracies per class on the validation set after linear probing. We name the created subsets as *CLIP-1K*, *DINOv2-1K*, *Sup-1K*, as well as their union. We compare the model-based subsets with subsets consisting of 1K random classes and ImageNet-1K.

D.4 Multi-label clustering benchmarks

Here, we investigate whether clustering methods capture additional labels, which is particularly useful for non-mutually exclusive datasets and large-scale datasets with multiple objects. Each image is associated with a set of labels, S_L, rather than just a single label.

For hierarchically structured datasets such as ImageNet21K, we add all semantic ancestors of the image's GT label to create S_L. This is motivated by the fact that, for example, an image of a guitarist (GT) can be classified as a musician or person, something

	Information		Probing	k-means			TEMI			SCANv2		
	# samples	# classes	ACC	ACC	NMI	BF1	ACC	NMI	BF1	ACC	NMI	BF1
ImageNet1K benchmarks												
Original	1.28M	1000	82.9	62.7	85.9	57.8	66.4	86.0	57.8	**69.8**	86.6	60.0
ODD	0.96M	1000	82.1	60.8	85.5	56.4	62.8	85.4	55.5	**65.8**	85.8	57.2
ImageNet21K benchmarks												
Imbal-2K	2.55M	2113	67.8	34.8	75.1	28.1	**36.0**	75.1	26.8	34.2	74.6	26.2
Centered-2K	2.34M	2113	68.5	35.2	75.2	28.5	**36.7**	75.2	27.2	34.5	74.6	26.5

Table 7.1: We report the ACC, normalized mutual information (NMI) and B-cubed F-1 (BF1) in % on imbalanced benchmarks. We use MAE-R [4] on ImageNet1K benchmarks and iBOT [64] on ImageNet21K benchmarks. Supervised results are marked in grey as they are not directly comparable to clustering methods.

that is not captured by existing clustering metrics. On ImageNet-1K, we adopt the multi-labeled validation set from [11], namely ImageNet-1K ReaL.

Multi-label metrics. Apart from the standard top 1 accuracy, we consider a weaker measure of accuracy where a predicted label is considered correct if it is contained in S_L. We refer to this as top $1 \rightarrow L$ accuracy. Further, we explore whether the top 5 cluster predictions contain useful information. To do this, we measure top 5 accuracy and top $5 \rightarrow L$ accuracy, where a sample is considered correctly classified if the top 5 predictions intersect with S_L.

To find the optimal mapping $f(.)$ from clusters to classes (Hungarian mapping [34]), we require a single prediction and label per sample. To this end, we take the cluster with the highest likelihood for TEMI and SCANv2 or the closest centroid for k-means as the prediction. To obtain a single label per image from S_L to estimate $f(.)$, we take the GT label when using ImageNet21k. For ImageNet-1K ReaL, we randomly sample a single label from the multiple GT labels provided for each image. We repeat this sampling 50 times and report the mean accuracies.

E Experimental Results

Implementation details. Following [42], we use the preprocessed winter 2021 version of ImageNet21K tree unless otherwise specified. We use the training data to develop the clustering method and evaluate on the validation split as in [3]. We conduct all the clustering experiments on an Nvidia A100 GPU with 40GB of VRAM. For TEMI and SCANv2, we use H=32 clustering heads and train for 50 and 25 epochs on ImageNet-1K and ImageNet21K, respectively. Both TEMI and SCANv2 are trained using AdamW [36] with a learning rate of $10^{-4}\frac{\text{batch size}}{256}$ and weight decay of 10^{-4}. Different from [3], we use a larger batch size (i.e. 4096 on ImageNet-1K instead of 512) for SCANv2 as we observed training instabilities and improved performance. For k-means clustering, we use cosine similarity as the distance metric. We report the clustering validation accuracy (ACC) using the Hungarian algorithm [34], the normalized mutual information (NMI), and the B-cubed F-1 score (BF1) [5]. Finally, we report the linear probing accuracy as a class separability measure and as an upper bound of clustering methods.

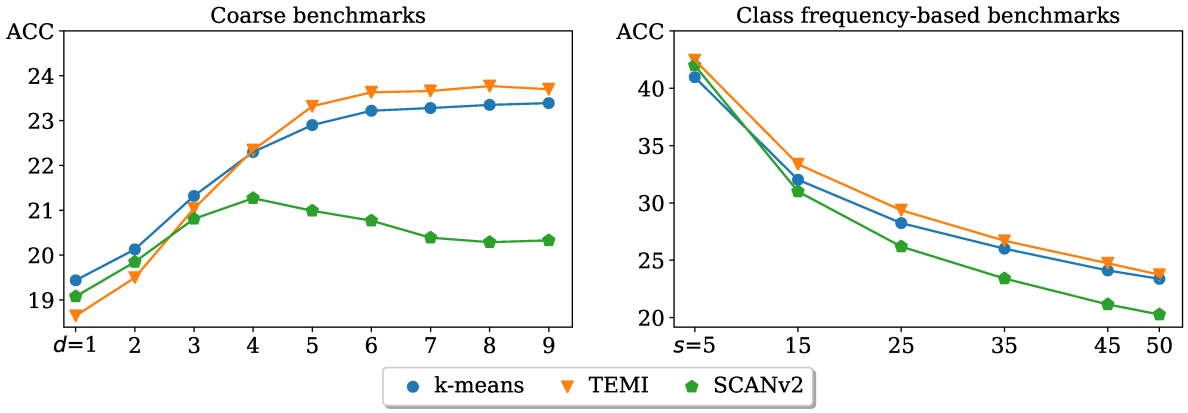


Figure 7.2: **Left:** Clustering accuracy in % (*y-axis*) versus maximum hierarchy depth d (*x-axis*) by mapping each ImageNet21K class to its semantic ancestor at depth d. **Right:** Clustering accuracy in % (*y-axis*) versus percentile s (*x-axis*) such that $m_s = [50-s, 50+s]$ based on ImageNet21K classes. We report the average ACC over three independent runs, with a maximum standard deviation of 0.1. Best viewed in color.

E.1 Impact of class imbalance

ImageNet-1K. We first verify the superiority of TEMI and SCANv2 on ImageNet-1K in Table 7.1 where we measure at least 5.9% and 10.0% relative ACC gain compared to k-means using MAE-R ViT-H [4]. With a well-tuned SCANv2 framework, we achieve a new state-of-the-art clustering ACC on ImageNet-1K of 69.79%. Notably, this accuracy surpasses a supervised GoogLeNet[2] [51] as a 68.3% top 1 ACC has been reported in [29]. We clarify that we increase the mini-batch size from 512 to 4096 to get the state-of-the-art ACC and explain why it is necessary in Section F.

Imbalanced benchmarks. For imbalanced datasets such as ImageNet-1K ODD and Imbalanced-2K, the ACC gains of TEMI and SCANv2 compared to k-means are reduced. We measure a relative ACC degradation of 3.2% for k-means, 5.4% for TEMI, and 5.7% for SCANv2. Similar behavior is observed on the ImageNet21K scale when comparing the Centered-2K versus Imbalanced-2K benchmarks. When reducing the percentage of odd class indices to $\leq 30\%$, k-means outperforms other methods (see appendix). When gradually scaling up the dataset size and the degree of imbalance simultaneously (Fig. 7.2, right), we observe the same trend: although TEMI consistently outperforms k-means, its relative ACC gain diminishes (from a maximum of 4.23% to 1.6% using all samples).

Based on the above, we state that a) k-means demonstrates suboptimal performance even on imbalanced benchmarks, and b) the degree of imbalance only partially accounts for the inferior performance of k-means observed in Table 7.1 and Fig. 7.2 as well as previous studies [6, 3, 16]. Another veiled aspect is that despite TEMI and SCANv2 enforcing class uniformity, they are applicable beyond the scope of ImageNet-1K, such as large-scale imbalanced datasets.

[2]https://paperswithcode.com/sota/image-classification-on-imagenet

	Information		k-means			TEMI			SCANv2		
	# samples	# classes	ACC	NMI	BF1	ACC	NMI	BF1	ACC	NMI	BF1
Leaf-based subsets											
WordNet	7.54M	7353	27.4	74.3	21.1	**28.1**	74.4	19.5	24.6	73.8	19.2
IN21K	8.37M	8153	26.7	74.1	20.4	**27.3**	74.2	20.3	23.6	73.5	20.0
Relabeled ImageNet21K benchmarks											
Leaf HZR	11.06M	8153	24.5	72.2	18.0	**24.6**	72.3	17.3	22.7	71.8	17.3
p-HZR	11.06M	10413	24.3	73.3	17.5	**24.5**	73.4	16.8	21.8	72.7	16.6
GT labels	11.06M	10450	23.4	73.1	17.1	**23.7**	73.2	16.5	20.3	72.4	16.5

Table 7.2: ACC and NMI of leaf subsets based on their respective semantic trees, fine-grained benchmarks based on HZR and ImageNet21K (GT).

E.2 Impact of class granularity

In Fig. 7.2 (left), we illustrate the methods' accuracies for different maximum hierarchy depths d. Interestingly, k-means outperforms TEMI for $d \leq 3$ with a maximum relative difference of 4.1% for $d = 1$. We argue that highly coarse labels are more likely to be captured by isotropic and normally distributed clusters assumed by k-means. Conversely, TEMI and SCANv2 rely on fixed NN pairs, which lead to retrieving similar-looking images to the query image [40], which hinders the coarse clustering performance for superclasses such as "animal".

Regarding the considered fine-grained benchmarks in Table 7.2, all clustering methods benefit similarly from purging mutually exclusive classes. The highest relative gain is observed in the subsets of WordNet leaf classes, where TEMI and k-means improve by 18.29% and 17.32% (relative ACC) compared to ImageNet21K. Overall, TEMI outperforms k-means on the majority of coarse and fine-grained ImageNet21K-based benchmarks, but the gains in Table 7.2 are marginal at the scale of 7.3K classes or more. Based on Fig. 7.2 and Table 7.2, we cannot identify any strong ACC discrepancy between clustering methods w.r.t. the class granularity.

E.3 Results from easy-to-classify benchmarks

All three model-based splits of 1K classes achieve high probing ACC ($> 95\%$) using iBOT (pretrained on ImageNet21K) as shown in Table 7.3. This indicates that the image features of easy-to-classify classes are well-separated, verifying [53]. Intriguingly, we measure a high discrepancy between k-means and deep learning methods independent of the model used to pick the top-1K classes (CLIP, DINOv2, Supervised). The reported discrepancy does not originate from the degree of imbalance. Compared to TEMI, k-means underperforms by a relative accuracy difference of at least 12.9%. We thus state that well-separated clustering benchmarks may have irregular class shapes [21] that cannot always be modeled with k-means. Methods such as TEMI and SCANv2 are more flexible by leveraging the structure of the feature space (through the NN) to capture irregular class shapes. The latter likely leads to a decision boundary of a more arbitrary form compared to k-means.

	Information	Probing	k-means		TEMI		SCANv2	
Subsets	# samples	ACC	ACC	NMI	ACC	NMI	ACC	NMI
(1) CLIP-1K	1.14M	95.8	55.1	84.5	**63.3**	86.3	62.9	85.8
(2) DINOv2-1K	1.13M	95.7	54.2	83.9	**61.2**	85.6	61.1	84.9
(3) Sup-1K	1.12M	96.0	55.4	84.6	62.5	86.4	**62.7**	85.8
Random 1K	1.06M	78.1	41.5	76.4	**44.1**	77.0	43.6	76.6
ImageNet-1K	1.28M	81.5	41.9	75.7	**47.3**	77.2	45.8	76.3
Union (1,2,3)	1.49M	94.6	52.1	83.2	**59.4**	84.9	58.0	84.2

Table 7.3: Accuracy and NMI on model-based ImageNet21K subsets based on different pre-trained models. We use the top-1K classes (highest ACC) of openCLIP ViT-G, DINOv2 ViT-g, and a supervised ViT-L (Sup). For the reported metrics, we use iBOT ViT-L pre-trained on ImageNet21K.

	Top1	Top1 →L	Top5	Top5 →L
ImageNet21K using iBOT ViT-L				
k-means	23.4	26.6	39.9	46.2
TEMI	23.7	27.1	42.7	48.8
SCANv2	20.3	22.9	31.8	37.8
Linear probing	47.3	51.3	76.4	79.8
ImageNet1K-ReaL using MAE-R ViT-H				
k-means	62.6	68.2	78.2	83.1
TEMI	67.3	73.1	82.0	86.8
SCANv2	70.2	75.9	84.6	89.4
Linear probing	81.1	88.2	94.4	97.3

Table 7.4: Multi-label clustering accuracies in % evaluated using the semantic ancestors on ImageNet21K and ImageNet-1K ReaL [11].

E.4 Multi-label clustering evaluations

As shown in Section E.4, the top 1 predictions capture relevant additional concepts (coarser labels, coexisting labels, or multiple objects of interest) as measured by top 1 →L. We measure a relative top 1→L ACC increase of over 13% (ImageNet21K) and 8.1% (ImageNet1K-ReaL) across clustering methods. Regarding the top 5 ACCs, the non-primary cluster predictions capture meaningful classes, corroborating with [11]. Unlike probing, where single-label supervision is provided during training, clustering methods achieve higher relevant ACC gains in top 5 and top 5 →L on both benchmarks.

F Discussion and Limitations

Class separability and clustering benchmarks. To further investigate the reported results in Table 7.3, we measure the GT alignment score [53], the silhouette score [44] and the Davies-Bouldin score [17]. All metrics highly correlate with the probing ACC,

specifically $R^2 > 0.97$, and point to the fact that model-based subsets are the most well-separated benchmarks. We highlight that model-based filtering is an automatic way to create new benchmarks without any explicit constraint on the class imbalance or granularity level.

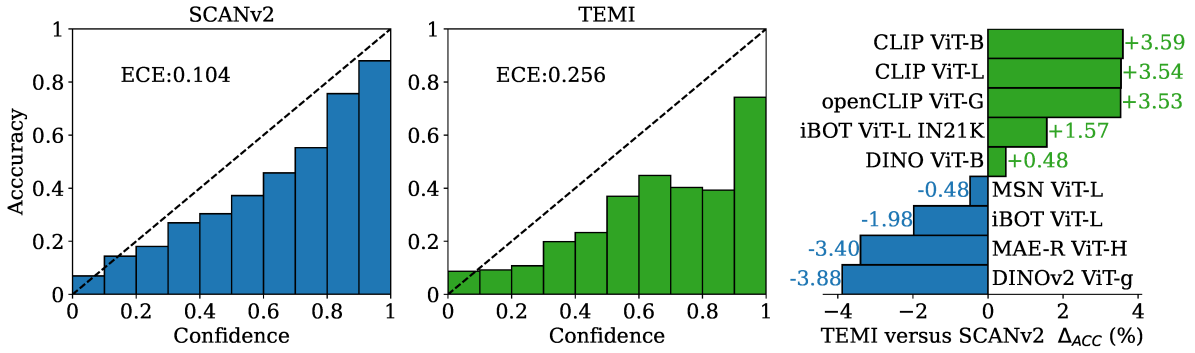


Figure 7.3: **Left and middle:** Calibration plots and expected calibration error (ECE ↓) for SCANv2 and TEMI on ImageNet-1K using MAE-R ViT-H [4]. **Right:** Clustering ACC difference of TEMI compare to SCANv2 on ImageNet-1K across various pre-trained feature extractors. IN21K refers to ImageNet21K, and CLIP ViT-B and ViT-L use the weights from [41].

Are we done with ImageNet-1K? No, but additional large-scale clustering benchmarks should be taken into account. Limitations and idiosyncrasies of ImageNet-1K have been pointed out in a series of classification-centered studies [11, 54]. From the perspective of clustering, ImageNet-1K is substantially closer to a random split of 1K ImageNet21K classes in terms of probing ACC (Table 7.3). Given that the training set of ImageNet-1K is balanced (Table 7.1), a non-inherent limitation of clustering, we state that it should not be used in isolation as the core large-scale benchmark.

Sensitivity of SCANv2 to the batch size. In contrast to [3], we identify that SCANv2 outperforms TEMI on ImageNet-1K (Table 7.1). Yet SCANv2 performs consistently inferior on the large-scale benchmarks (Table 7.2), which is partially attributed to its sensitivity to the batch size B. In particular, the entropy term in Eq. (7.2) requires B to be sufficiently larger than the number of clusters C to approximate the cluster distribution $p(c)$. The latter requires a lot of extra memory as the space complexity scales linearly $O(C \times H)$ w.r.t. C and H (number of heads). This computational limitation becomes quite significant on large-scale benchmarks, as shown in Table 7.2.

Calibration analysis of TEMI and SCANv2 Unlike k-means, TEMI and SCANv2 provide a measure of confidence as measured by the maximum softmax probability (MSP) [26]. In Fig. 7.3 (left and middle), we show that even though TEMI yields more discriminative predictions (91.9% versus 80.2% mean MSP), the predictions of SCANv2 are better calibrated on ImageNet-1K. This is attributed to the entropy regularization in Eq. (7.2), which explicitly down-weights images with high confidence in the mini-batch. The confidence of well-calibrated clustering approaches can be leveraged in many applications, such as dataset pruning and semi-automatic data annotation.

Dependence on feature extractors. A limitation of this study is the dependence of feature-based clustering methods on the learned representations. Both TEMI and

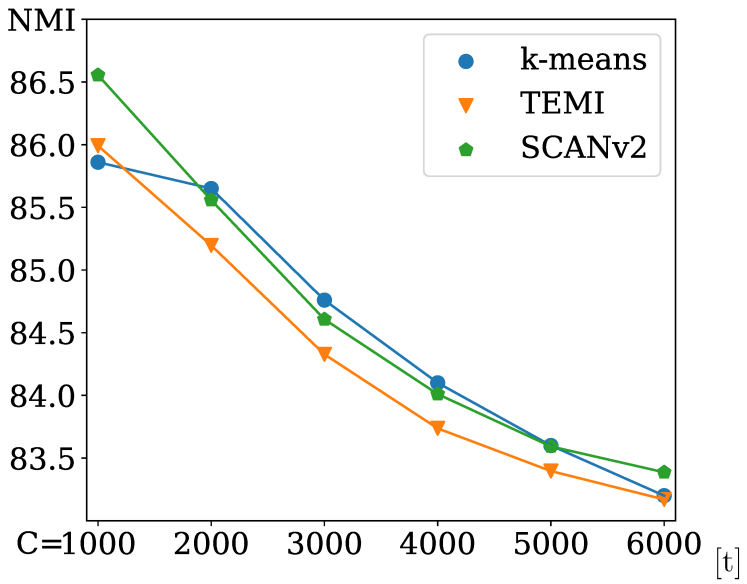


Figure 7.4: Normalized mutual information (NMI ↑, *y-axis*) versus the number of clusters C (*x-axis*) on ImageNet-1K.

SCANv2 consistently outperform k-means across feature extractors on ImageNet-1K. However, as shown in Fig. 7.3 (right), no method is consistently superior when varying pre-trained feature extractors. Based on that, we believe there is room for further improvements.

Overclustering on ImageNet-1K. Estimating the clustering ACC requires prior knowledge of the number of GT classes. In Fig. 7.4, we examine the impact of increasing C using the NMI metric. We observe that NMI progressively diminishes as C surpasses the GT classes for all methods. In future work, the hyperparameters β and α in TEMI and SCAN could be adjusted to downweight the class uniformity constraint as C increases. This further underscores the challenge of large-scale image clustering, particularly in the absence of a reliable estimate of the number of GT classes, which remains an open problem.

G Conclusion

In this paper, a comprehensive experimental study of feature-based clustering approaches and large-scale benchmarks was conducted. New clustering benchmarks based on ImageNet21K were created, focusing on the effects of individual factors, such as class imbalance, class granularity, and easily separable classes. Apart from understanding k-means, the introduced benchmarks can aid future practitioners in holistically evaluating large-scale foundational models beyond ImageNet. We believe that the new benchmarks will facilitate the comparison of new clustering approaches and vision-language model evaluation.

References

[1] Amro Abbas, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari S Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. *arXiv preprint arXiv:2401.04578*, 2024.

[2] Nikolas Adaloglou, Tim Kaiser, Felix Michels, and Markus Kollmann. Rethinking cluster-conditioned diffusion models. *arXiv preprint arXiv:2403.00570*, 2024.

[3] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[4] Benedikt Alkin, Lukas Miklautz, Sepp Hochreiter, and Johannes Brandstetter. Mim-refiner: A contrastive learning boost from intermediate pre-trained representations. *arXiv preprint arXiv:2402.10093*, 2024.

[5] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12:461–486, 2009.

[6] Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022.

[7] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 456–473. Springer, 2022.

[8] Fan Bao, Chongxuan Li, Jiacheng Sun, and Jun Zhu. Why are conditional generative models better than unconditional ones? In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[9] Björn Barz and Joachim Denzler. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, 2020.

[10] Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):5418, 2020.

[11] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[13] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[14] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] Tianzhe Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin David Haeffele, Rene Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-1(2):224–227, 1979.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Tianjiao Ding, Shengbang Tong, Kwan Ho Ryan Chan, Xili Dai, Yi Ma, and Benjamin D Haeffele. Unsupervised manifold linearizing and clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5450–5461, 2023.

[20] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[21] Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

[22] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[24] Wengang Guo, Kaiyan Lin, and Wei Ye. Deep embedded k-means clustering. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 686–694. IEEE, 2021.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[27] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[28] Laura Hollink, Aysenur Bilgin, and Jacco Van Ossenbruggen. Is it a fruit, an apple or a granny smith? predicting the basic level in a concept hierarchy. *arXiv preprint arXiv:1910.12619*, 2019.

[29] Forrest N Iandola, Matthew W Moskewicz, Khalid Ashraf, and Kurt Keutzer. Firecaffe: near-linear acceleration of deep neural network training on compute clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2592–2600, 2016.

[30] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[31] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[32] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[34] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[35] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[37] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[38] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[39] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023.

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[42] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[43] Meitar Ronen, Shahaf E Finder, and Oren Freifeld. Deepdpm: Deep clustering with an unknown number of clusters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9861–9870, 2022.

[44] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[47] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[48] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5484–5494, 2023.

[49] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[50] Armstrong Subero and Armstrong Subero. Tree data structures. *Codeless Data Structures and Algorithms: Learn DSA Without Writing a Single Line of Code*, pages 31–40, 2020.

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[52] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[53] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[54] Shixian Wen, Amanda Sofie Rios, Kiran Lekkala, and Laurent Itti. What can we learn from misclassified imagenet images? *arXiv preprint arXiv:2201.08098*, 2022.

[55] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[56] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[57] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[58] Xiao Xu, Shifei Ding, Yanru Wang, Lijuan Wang, and Weikuan Jia. A fast density peaks clustering algorithm with sparse search. *Information Sciences*, 554:61–83, 2021.

[59] Nianwen Xue. Steven bird, evan klein and edward loper. natural language processing with python. o'reilly media, inc. 2009. isbn: 978-0-596-51649-9. *Natural Language Engineering*, 17(3):419–424, 2011.

[60] William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-OOD: Deciphering modern out-of-distribution detection algorithms. In *The Twelfth International Conference on Learning Representations*, 2024.

[61] Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring hierarchical graph representation for large-scale zero-shot image classification. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022.

[62] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

[63] Zhenhua Yu, Yanghao Yan, Fan Deng, Fei Zhang, and Zhiwu Li. An efficient density peak cluster algorithm for improving policy evaluation performance. *Scientific Reports*, 12(1):5000, 2022.

[64] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

[65] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

Appendix

App. I Linear separability metrics.

In Fig. 7.5, we observe that all metrics (linear probing ACC, $\mathcal{L}_{\mathsf{align}}$, Davies-Bouldin score, Silhouette) of class separability are strongly correlated ($R^2 > 0.98$). First, the model-based ImageNet-21K benchmarks are well-clustered in the feature space of iBOT. Second, ImageNet-1K is much closer to a random split of 1K classes from ImageNet21K. $\mathcal{L}_{\mathsf{align}}$ is the alignment of features that share the GT same label (p_{pos}) as $\mathcal{L}_{\mathsf{align}}(g) = \mathbb{E}_{(x_1,x_2)\sim p_{\mathsf{pos}}}\|g(x_1) - g(x_2)\|_2^2$ [53]. [53] have used augmentations of the same image as positive pairs (p_{pos}) while we use true positive nearest neighbors, meaning images that share the same labels.

App. II CLIP zero-shot classification and calibration analysis on ImageNet21K

We perform a calibration analysis [23] on the zero-shot classification of openCLIP ViT-G on ImageNet21K) in Fig. 7.6. We find that the baseline zero-shot ACC using all 10450 classes of 36.02% can be increased to 55.10% when considering the subset of class labels from the same hierarchy. When further restricting the candidate labels to the semantic siblings of the GT class, if available, we measure an accuracy of 68.35%. Siblings refer to all class labels that share the same semantic ancestor as the GT. When semantic siblings are not available, we use the labels from the same hierarchy. In parallel, we find that CLIP is better calibrated when restricting the number of candidate labels to siblings,

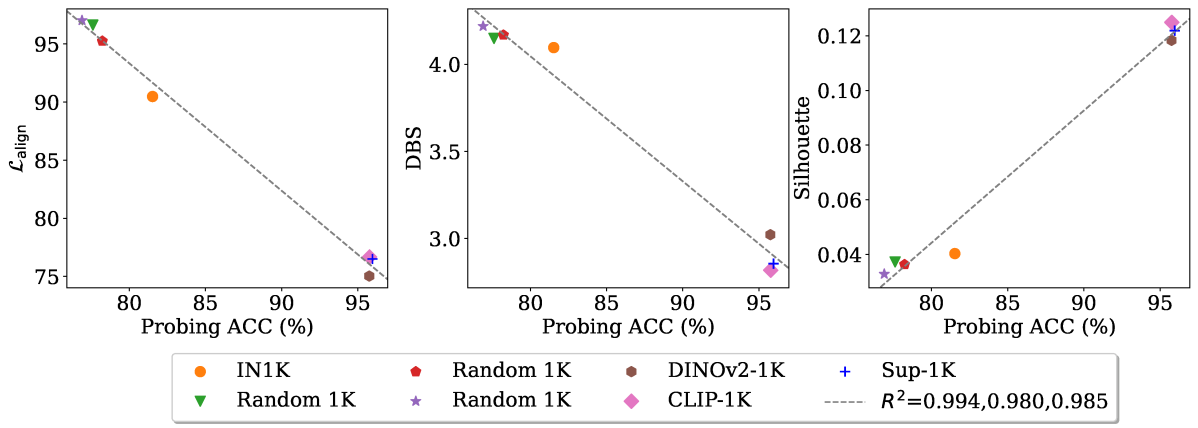


Figure 7.5: We measure the dependence of the linear probing accuracy *x-axis* versus the alignment score $\mathcal{L}_{\text{align}}$ (**left**), the Davies-Bouldin score (DBS) [17] (**middle**) and the Sihlouette score (**right**) using iBOT ViT-L trained on ImageNet21K. R^2 is the coefficient of determination. Best viewed in color.

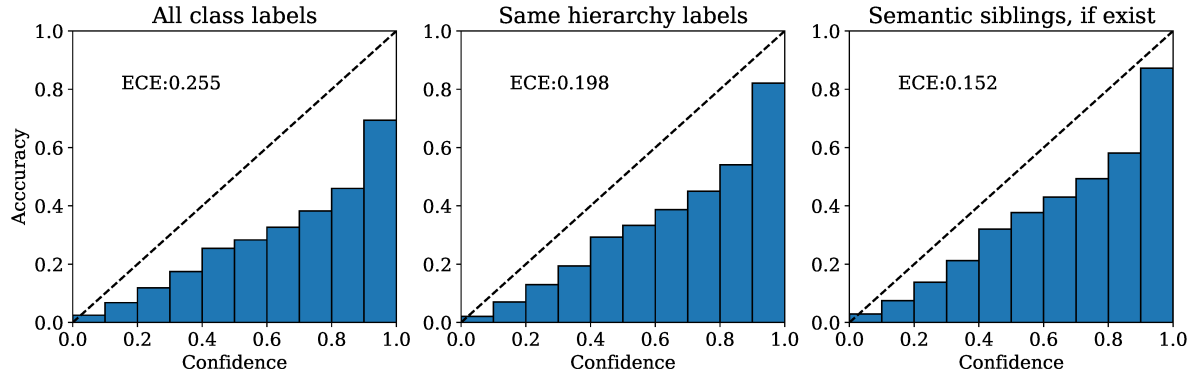


Figure 7.6: **Left:** Zero-shot classification using OpenCLIP with all 10450 class names on ImageNet21K (36.02% mean accuracy). **Middle:** We restrict the candidate labels for zero-shot classification to the labels of the same hierarchy (55.10% mean accuracy). **Right:** We further restrict the candidate labels to the semantic siblings based on the semantic tree (68.35% mean accuracy). When there are no semantic siblings, we use the classes from the same hierarchy. ECE stands for expected calibration error.

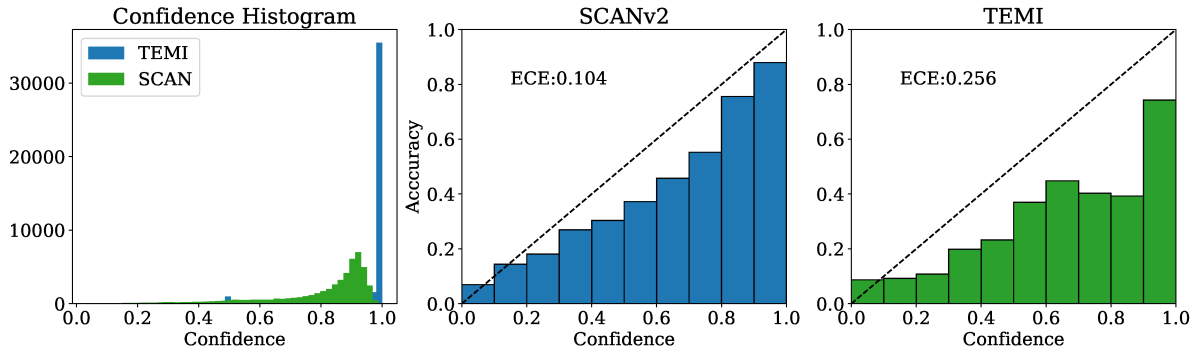


Figure 7.7: **Left:** Histogram of maximum softmax probabilities (confidences) on ImageNet-1K. **Middle, right:** Calibration plots for SCANv2 and TEMI. ECE stands for expected calibration error.

as measured by a relative decrease of 40% in the expected calibration error (ECE) [23]. We measure a mean confidence (maximum softmax probability [26] on the image-text similarities) of 83.44% versus 61.4% (baseline) using the semantic siblings.

App. III Confidence histogram of deep clustering methods on ImageNet-1K

While we have shown in the main text that SCANv2 is better calibrated than TEMI, their confidence histogram looks significantly different on ImageNet-1K, as illustrated in Fig. 7.7. TEMI produces more discriminative predictions, likely due to the pointwise mutual information objective. However, this makes its predictions less calibrated. We leave this finding for future investigation.

App. IV Varying the percentage of odd class indices on ImageNet-1K training set

In [19], the authors have only investigated the impact of halving the number of samples from odd class indices in balanced classification datasets. Here, we use varying percentages p of the samples from odd class indices. This allows us to quantify the impact of stronger artificial class imbalances on ImageNet-1K as depicted in Fig. 7.8. We find that feature-based k-means outperforms deep clustering methods for $p \leq 30\%$ for both iBOT ViT-L and MAE-R ViT-H pre-trained on ImageNet-1K. This further supports our claim that feature-based k-means is unfairly treated when solely evaluated in balanced classification datasets.

App. V Label reassessment using p-HZR.

Limitations of p-HZR. Regarding p-HZR, we identify many semantic links in WordNet that do not always align with sensible visual-based concepts and likely limit the effectiveness of p-HZR. Examples include person→father-in-law, animal→adult. WordNet class names that are not sufficiently descriptive in terms of visual information likely

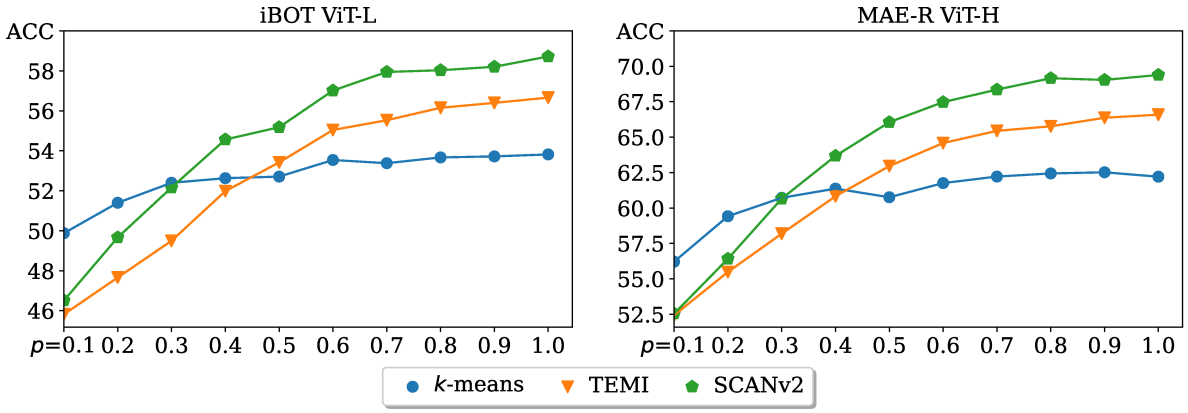


Figure 7.8: Clustering accuracy in % (*y-axis*) versus percentage of p remaining samples for odd class indices using the iBOT ViT-L (*right*) and MAE-R ViT-H (*left*) feature extractors. We use ImageNet-1K to derive the odd subsets with different p.

decrease the probability of reassessing a sample to a more fine-grained label using p-HZR. For example, the GT class "outdoor game" in the tree path {athletic game → outdoor game → field game → football → soccer} keeps images depicting football games to their GT label "athletic game". That is why we hypothesize that a more visually aligned tree would improve automatic tree-based label reassessment. Nonetheless, deriving a single gold standard label [28, 11] (the granularity level at which humans assign the GT label quicker and with greater accuracy) requires humans in the loop. We generally find multi-label clustering evaluations more informative on large-scale datasets, something that has not been sufficiently investigated by previous works [52].

Linear separability of p-HZR on ImageNet21K. An increase in linear separability w.r.t the refined labels compared to the GT suggests that refined labels align better with the learned features. Surprisingly, p-HZR yields a relative probing ACC gain of 6.8% while keeping 99.6% of the classes. A similar probing ACC was measured for the coarse benchmark for $d = 4$ (80.2% of the classes).

Additional information regarding p-HZR. The proposed method for reannotating ImageNet21K (p-HZR) only reduces the number of classes by 37 (train and validation), which requires excluding only 484 training samples ($< 0.01\%$ of the samples). Overall, we are able to refine $\approx 42\%$ of the training GT labels that are not leaf classes, which translates to more than 10% of ImageNet21K.

App. VI Zero-shot classification using CLIP on ImageNet21K

During all our zero-shot experiments with CLIP (openCLIP ViT-G), we first match the WordNet's unique identifier to one of the words in its set of synonyms, typically called lemmas, such as n00007846→[person, individual, someone, soul]. To this end, we perform zero-shot classification across lemmas for all images of the same class and subsequently apply majority voting. This step effectively reflects the linguistic biases of the pre-training data used to train the vision-language model, namely LAION-2B [46]. For example, CLIP prefers "artefact" instead of "artifact". We use a subset of 7 out of the 80 prompts initially proposed in [41], namely "itap of a {label}", "a bad photo of

the {label}", "an origami {label}", "a photo of the large {label}", "a {label} in a video game", "art of the {label}", "a photo of the small {label}", based on subsequent prompt engineering analysis of the same authors[3].

App. VII Additional information

Linear probing pipeline and hyperparameters. For the linear probing experiments, we perform a grid search on a set of learning rates and weight decays. We train the linear layer with stochastic gradient descent for 100 epochs without augmentations and report the highest validation accuracy value obtained. We use the values of $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$ and $\{0, 10^{-3}, 10^{-5}\}$ for the learning rate and weight decay, respectively.

Clustering algorithm hyperparameters Both TEMI and SCANv2 are trained using AdamW [36] with a learning rate of $10^{-4}\frac{\text{batch size}}{256}$ and weight decay of 10^{-4}. Furthermore, both methods use 32 heads, which are MLPs with a single hidden layer of size 512 and GELU [27] activations. The algorithm-specific hyperparameters described in Section C are $\beta = 0.6$ for TEMI and $\alpha = 5$ for SCANv2. We apply temperature scaling with $\tau = 0.1$ before the softmax for both SCANv2 and TEMI.

Dataset information for the considered ImageNet21K subsets Table 7.5 depicts all dataset-related information for the considered frequency-based benchmarks and coarse datasets. We highlight that all datasets have been created based on the preprocessed ImageNet21K winter 2021 version from [42].

Reassessed ImageNet21K samples from p-HZR. Below, we have collected some reassessed ImageNet21K samples using p-HZR to show the effectiveness of not always refining the GT label to a more fine-grained level.

[3] `https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb`

Table 7.5: We use MAE-R [4] on ImageNet1K benchmarks and iBOT [64] on ImageNet21K benchmarks. Bold indicates the highest clustering accuracy, and supervised results are marked in grey.

	Information		Accuracy (%)			
	# samples	# classes	Probing	k-means	TEMI	SCANv2
ImageNet21K frequency-based benchmarks						
Tiny ($m_s = [45, 55]$)	1.17M	1060	76.98	40.96	**42.46**	41.99
Small ($m_s = [35, 65]$)	3.46M	3146	63.71	32.02	**33.38**	30.99
Base ($m_s = [25, 75]$)	5.67M	5230	56.53	28.24	**29.36**	26.19
Large ($m_s = [15, 85]$)	7.79M	7326	52.04	26.01	**26.70**	23.40
Huge ($m_s = [5, 95]$)	9.88M	9405	48.57	24.11	**24.74**	21.15
Coarse ImageNet21K benchmarks						
Coarse $d = 1$	11.06M	3987	51.56	**19.44**	18.65	19.08
Coarse $d = 2$	11.06M	5086	51.97	**20.13**	19.50	19.85
Coarse $d = 3$	11.06M	6772	51.48	**21.32**	21.04	20.81
Coarse $d = 4$	11.06M	8386	49.99	22.30	**22.34**	21.27
Coarse $d = 5$	11.06M	9429	48.74	22.90	**23.32**	20.99
Coarse $d = 6$	11.06M	9997	47.78	23.22	**23.63**	20.77
Coarse $d = 7$	11.06M	10241	47.19	23.28	**23.66**	20.39
Coarse $d = 8$	11.06M	10351	46.91	23.35	**23.77**	20.29
ImageNet21K	11.06M	10450	46.72	23.38	**23.74**	20.27

Figure 7.9: Reassessed ImageNet21K samples using p-HZR. Each subplot's title shows the originally annotated ground truth label. Rows display the reassessed labels. Samples were not reannotated if the reassessed label matched the ground truth label.

8. ClusterMine: Robust Label-Free Visual Out-Of-Distribution Detection from Text Corpora

Outline

A Introduction 1
B Related work 2
C OOD detection methods using CLIP 3
D Experimental evaluation 6
E Future work and conclusion 13
References 15

Bibliographic Information

Adaloglou, N., Petrusheva*, D., Asker*, M., Michels, F. , and Kollmann, M. (2026). ClusterMine: Robust Label-Free Visual Out-Of-Distribution Detection via Concept Mining from Text Corpora. In 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE.

Corresponding author: *adaloglo@hhu.de*

Copyright Notice

This chapter is based on the author's paper published in WACV 2026, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2026. The author retains the copyright to this work. The copyright was adjusted according to the guidelines of the publisher.

Acknowledgements

The author would like to thank the co-authors, Mohamed Asker and Diana Petrusheva, for prototyping and conducting the preliminary experiments with the new label mining approaches.

Individual Contributions

The author identified a critical gap in the literature: existing approaches leveraging vision-language models for out-of-distribution (OOD) detection assume prior knowledge of the dataset's class label names. To address this limitation, the author developed two simple yet novel methods for out-of-distribution detection. The prototyping phase was conducted together with Mohamed Asker and Diana Petrusheva. Apart from large-scale OOD detection benchmarks, the author extended the scope of the analysis to include robustness to in-distribution shifts. The author also conducted a scalability analysis concerning the model size. Finally, the author carried out the majority of the experimental work and wrote most of the manuscript.

Abstract

Large-scale visual out-of-distribution (OOD) detection has witnessed remarkable progress by leveraging vision-language models such as CLIP. However, a significant limitation of current methods is their reliance on a pre-defined set of in-distribution (ID) ground-truth label names (positives). These fixed label names can be unavailable, unreliable at scale, or become less relevant due to in-distribution shifts after deployment. Towards truly unsupervised OOD detection, we utilize widely available text corpora for positive label mining, bypassing the need for positives. In this paper, we utilize widely available text corpora for positive label mining under a general concept mining paradigm. Within this framework, we propose ClusterMine, a novel positive label mining method. ClusterMine is the first method to achieve state-of-the-art OOD detection performance without access to positive labels. It extracts positive concepts from a large text corpus by combining visual-only sample consistency (via clustering) and zero-shot image-text consistency. Our experimental study reveals that ClusterMine is scalable across a plethora of CLIP models and achieves state-of-the-art robustness to covariate in-distribution shifts. The code is available at `https://github.com/HHU-MMBS/clustermine_wacv_official`.

A Introduction

Given a set of training images comprising the in-distribution (ID), visual out-of-distribution (OOD) detection aims to identify images sampled from a shifted distribution while having access only to the ID [66, 2]. For each ID, there exists a predefined set of semantic categories $\mathcal{Y}_{\text{real}}$ that can be assigned to each sample [68].

Recently, distribution shifts have been categorized into semantic shifts ($\mathcal{Y}_{\text{OOD}} \cap \mathcal{Y}_{\text{real}} = \varnothing$) and non-semantic shifts, also known as covariate shifts. A covariate-shifted ID pair (x, y) preserves its label $y \in \mathcal{Y}_{\text{real}}$ while x undergoes a distribution shift [5, 69]. In the context of natural image recognition, examples of covariate shifts can arise from camera variations, background changes, or style shifts [24, 20, 58].

Motivation. In practice, the ID images exhibit varying levels of class specificity and overlapping concepts [51, 29, 42, 4, 1]. While $\mathcal{Y}_{\text{GT}}$ often being a good proxy of $\mathcal{Y}_{\text{real}}$ in small data regimes, GT label names at scale can be incomplete, unreliable, or arbitrarily defined. Deriving a comprehensive set of visual concepts that fully characterizes the ID is challenging [57]. For instance, automated curation methods, such as hashtags [52, 40], are frequently underdescriptive or inconsistent with natural language [6, 63]. Fixed label names can bottleneck performance when class semantics shift due to in-distribution drifts, a common occurrence after deployment. To overcome these challenges, our work proposes to extract ID concepts directly from large text corpora that are aligned with the ID images.

Detecting covariate shifts. An ideal OOD detector should be sensitive to semantic shifts and, at the same time, robust to covariate shifts [64]. Misclassified covariate-shifted samples can undermine the generalization capability of an ID classifier. When these samples are flagged as OOD, they render the classifier's predictions unreliable [68]. Yang et al. [66] have recently demonstrated that existing supervised OOD detection algorithms are susceptible to covariate shifts. To date, the OOD robustness to covariate shifts has not been sufficiently explored [63, 16, 67].

Scalability and vision-language (VL) models. In parallel, VL models have been established for large-scale OOD detection benchmarks [46, 19, 43, 15, 2]. However, the majority of literature comparisons [44, 70, 61] are conducted with small-sized models relative to the pretraining dataset, such as CLIP ViT-B [13, 46]. Smaller-sized models bottleneck performance when other scaling factors remain constant [34, 11]. This raises questions whether the recently developed methods are scalable with respect to model size. Plain image-text similarities have shown non-optimal performance as the model size increases [2]. Last but not least, current VL methods for OOD detection *assume access to* $\mathcal{Y}_{\text{GT}}$, which simplifies the task at hand.

Questionable benchmarks. Another point of criticism concerns the adopted large-scale OOD benchmarks. Huang et al. [30] introduced a quadruplet of OOD benchmarks for ImageNet with unknown levels of semantic overlap and limited diversity. This benchmark suite remains widely adopted to date [33, 61], despite recent studies raising concerns about its validity [66]. For instance, Bitterwolf et al. [9] estimate a semantic overlap of almost 60% in the Places OOD benchmark compared to ImageNet [12]. Several newer datasets have been carefully designed to minimize semantic overlap with the ID [9].

The above factors - *reliance of VL models on ID label names, covariate shifts, scalability, benchmarks* - OOD detection methods may have begun to overfit the idiosyncrasies

of certain models and benchmarks, mirroring concurrent trends observed in image classification [7]. In other words, it is unclear how robust the recent advancements in large-scale OOD detection using CLIP models are.

In this paper, we present a general framework for extracting relevant concepts from a large text corpus that encompasses existing methods. Moreover, we develop the first positive label name mining method, ClusterMine, which extracts high-quality ID-related concepts from the corpus. A systematic large-scale OOD benchmarking reveals that ClusterMine i) scales with model size and performance, ii) achieves state-of-the-art OOD detection AUROC on most benchmarks *without access to ground truth ID label names*, and iii) is robust against most ID shifts on ImageNet, iv) outperforms state-of-the-art approaches in near-OOD benchmarks, such as AdaNeg [70] that impose additional requirements (i.e. sequential access to OOD samples).

B Related work

Supervised visual OOD detection methods. Supervised OOD detection involves training or fine-tuning a classifier on labeled ID samples [38, 32, 60, 35]. Posthoc detectors then typically compute an OOD score such as the maximum softmax probability (MSP) [25]. Alternative training-time modifications include regularizations by incorporating OOD data [39, 62, 38], auxiliary samples [26], or synthetic outliers [14]. However, when benchmarked rigorously and at scale, Yang et al. [66] demonstrate that no OOD detection score [39, 22, 71, 53] consistently outperforms MSP when applied to a supervised classifier, which has shifted the focus to VL models.

OOD detection using CLIP. Fort et al. [17] utilize CLIP by combining ID and OOD candidate class names for zero-shot inference. Nonetheless, prior knowledge of OOD class names is rarely available in real-world scenarios. Subsequently, Ming et al. [43] show that simply using MSP on image-text similarities, commonly referred to as *maximum concept matching (MCM)*, is sufficient for OOD detection using CLIP. Here, the text representations are computed from the ID label names. In [15], the authors introduce an additional text decoder on top of CLIP's image encoder to generate candidate labels, which can result in overlapping labels with ID data. Using artificially generated OOD data, [54, 36] finetune CLIP to learn a decision boundary between ID and OOD data.

Recently, Galil et al. [19] demonstrated that CLIP can function as a capable zero-shot detector without fine-tuning. To date, the only approach that showed promising scaling behavior is by Adaloglou et al. [2], who propose a two-step approach to train a head using pseudo-labels derived from zero-shot inference. Additionally, CLIP provides a flexible way to define which classes are considered ID during inference [19]. This aspect remains relatively unexplored, as prior methods rely on the predefined set of ground-truth ID class names, which may not always be optimal.

Incorporating negative concepts into CLIP. A promising direction using CLIP models is the integration of negative prompts or negative labels [17, 61, 37, 33, 70]. Negative prompt learning methods introduce a learnable "negative" prompt along with a dedicated "negative" text encoder to capture negation semantics in images [61]. Such an approach requires auxiliary data and the training of additional text-based components. Li et al. [37] circumvent these limitations by learning negative text prompts using CLIP's existing text encoder, enabling the model to specialize in capturing negative semantics

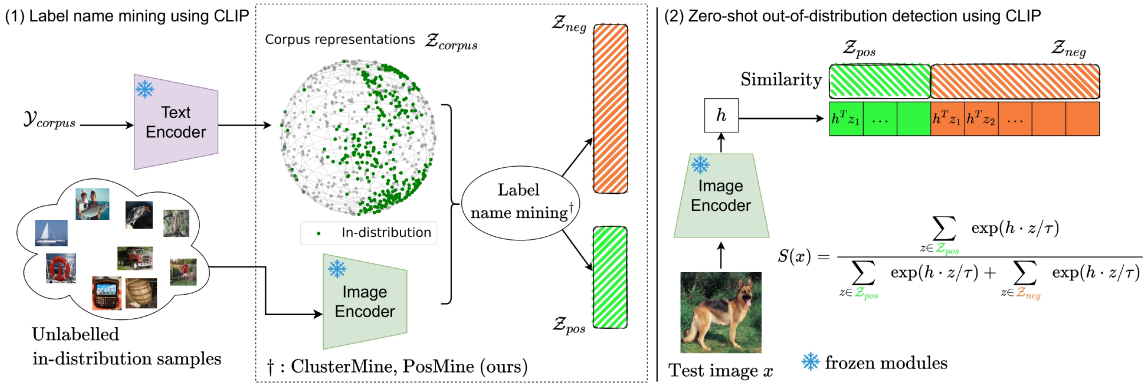


Figure 8.1: **An overview of the label mining framework for OOD detection using CLIP**. Given a text corpus $\mathcal{Y}_{\text{corpus}}$ and its feature representation $\mathcal{Z}_{\text{corpus}}$, ClusterMine and PosMine aim to extract in-distribution-related class names $\mathcal{Y}_{\text{pos}}$ in the shared vision-language space of CLIP. $\mathcal{Z}_{\text{neg}}$ can be either realized as the non-overlapping elements of $\mathcal{Y}_{\text{pos}}$ and $\mathcal{Y}_{\text{corpus}}$, or as the most dissimilar text representations from $\mathcal{Z}_{\text{pos}}$ (negative label mining). The OOD detection score is $S(x)$. Best viewed in color.

relative to ID classes.

Negative labels refer to adding class names during zero-shot inference. The additional class names likely capture OOD-related classes. A naive approach is to use the $\mathcal{Y}_{\text{OOD}}$, as employed by Fort et al.[17]. Our work is more closely related to NegLabel [33], a training-free method that identifies negative labels based on their dissimilarity from $\mathcal{Y}_{\text{GT}}$ using a text corpus $\mathcal{Y}_{\text{corpus}}$. More recently, [70] developed adaptive variants of NegLabel that dynamically update the negative concepts at test time as more OOD test images become available.

OOD detection robustness to covariate shifts. Distribution shifts are typically categorized into covariate and semantic shifts [65]. Covariate shift is usually associated with model calibration [55, 21, 45, 10]. Yang et al. [65] coined the term "full-spectrum detection" in the context of small-scale benchmarks. Full-spectrum simultaneously considers semantic shifts (OOD detection) and covariate shifts (also known as OOD robustness or OOD generalization [24, 23, 55]) within the OOD evaluation pipeline. Modern supervised OOD detection algorithms remain quite susceptible to non-semantic covariate shifts [64, 68]. By extracting negative concepts from a text corpus, NegLabel [33] reports notable improvements in robustness to covariate shifts compared to standard MCM. This suggests that CLIP can be tailored to the ID by selecting suitable negative and positive textual inputs.

C OOD detection methods using CLIP

Notation. We define a sufficiently large text corpus of possible label names $\mathcal{Y}_{\text{corpus}} = \{y_1, y_2, \ldots, y_N\}$ with cardinality $|\mathcal{Y}_{\text{corpus}}| = N$. By mining ID images, we extract from the corpus positive and negative label sets, $\mathcal{Y}_{\text{pos}}, \mathcal{Y}_{\text{neg}}$, with $\mathcal{Y}_{\text{pos}} \cap \mathcal{Y}_{\text{neg}} = \varnothing$ and $\mathcal{Y}_{\text{pos}} \cup \mathcal{Y}_{\text{neg}} \subseteq \mathcal{Y}_{\text{corpus}}$. The set $\mathcal{Y}_{\text{corpus}}$ is designed such that $\mathcal{Y}_{\text{pos}}$ is strongly related to the real semantic categories of the ID, whereas $\mathcal{Y}_{\text{neg}}$ is not. We use the text encoder of CLIP to generate the

representations $\mathcal{Z}_{pos}, \mathcal{Z}_{neg}$, which correspond to $\mathcal{Y}_{\text{pos}}, \mathcal{Y}_{\text{neg}}$, respectively. We demonstrate the general framework in Fig. 8.1, which is described below.

C.1 A general OOD detection framework for CLIP

Vision-language models can leverage a secondary set of "negative" concepts $\mathcal{Y}_{\text{neg}}$ that are unrelated to the categories of the ID. Given an image x and its representation $h = g(x)$ from the CLIP image encoder $g(.)$, we can define an OOD score by

$$S(x) = \frac{\sum_{z \in \mathcal{Z}_{pos}} \exp(h \cdot z/\tau)}{\sum_{z \in \mathcal{Z}_{pos}} \exp(h \cdot z/\tau) + \sum_{z \in \mathcal{Z}_{neg}} \exp(h \cdot z/\tau)}, \tag{8.1}$$

with temperature parameter $\tau > 0$. Both image and text representations are unit vectors. The particular choice of $\mathcal{Y}_{\text{pos}}, \mathcal{Y}_{\text{neg}}$ depends on the method. For instance, Fort et al. [17] use Eq. (8.1) with $\mathcal{Y}_{\text{pos}} = \mathcal{Y}_{\text{GT}}$ and $\mathcal{Y}_{\text{neg}} = \mathcal{Y}_{\text{OOD}}$, which greatly simplifies the task. NegLabel [33] also assumes $\mathcal{Y}_{\text{pos}} = \mathcal{Y}_{\text{GT}}$ and extracts $\mathcal{Y}_{\text{neg}}$ from a text corpus $\mathcal{Y}_{\text{corpus}}$ using negative label mining.

Negative label mining [33] first computes the cosine similarities between $\mathcal{Z}_{corpus}$ and $\mathcal{Z}_{pos}$. Excluding overlapping concepts from the corpus,

$$\mathcal{Y}_{\text{neg}} = \mathcal{Y}_{\text{corpus}} \setminus \mathcal{Y}_{\text{pos}}, \tag{8.2}$$

the $K \leq |\mathcal{Y}_{\text{neg}}|$ most dissimilar text representations from $\mathcal{Y}_{\text{neg}}$ are considered. In [33], the authors use the percentile distance, i.e. a 95% percentile instead of the minimum. For the edge case of $K = |\mathcal{Y}_{\text{neg}}|$, no pruning is applied.

Negative grouping and dynamic negative mining. By splitting $\mathcal{Y}_{\text{neg}}$ into randomly sampled groups, Eq. (8.1) can be applied per group [33]. The final score is the average across groups. We did not observe any performance gains from using this strategy; therefore, it is not employed in this work. When OOD samples appear sequentially, the choice of negative concepts can be adjusted at test time [70]. Such strategies can be integrated into the presented framework, and we leave them for future work.

C.2 Positive label mining methods

Different from previous works [17, 33], positive label mining aims to design OOD approaches using CLIP models that do not depend on prior knowledge of $\mathcal{Y}_{\text{GT}}$. At test time, Eq. (8.1) is applied after estimating $\mathcal{Y}_{\text{pos}}, \mathcal{Y}_{\text{neg}}$ from $\mathcal{Y}_{\text{corpus}}$. Other post-hoc training scores, such as pseudo-label probing [2], achieved inferior performance.

We first test whether "naive" positive mining from $\mathcal{Y}_{\text{corpus}}$ using zero-shot inference is comparable to existing methods that rely on $\mathcal{Y}_{\text{GT}}$. Afterwards, we consider a label name as ID ($y \in \mathcal{Y}_{\text{pos}}$) if at least M training samples are assigned to this particular class and $\mathcal{Y}_{\text{neg}} = \mathcal{Y}_{\text{corpus}} \setminus \mathcal{Y}_{\text{pos}}$. Alternative formulations of PosMine that explicitly control the cardinality of $\mathcal{Y}_{\text{pos}}$ can be realized. While results are affirmative (Table 8.1), selecting M becomes challenging without an OOD validation set. In addition, PosMine solely checks for image-text similarities to derive $\mathcal{Y}_{\text{pos}}$.

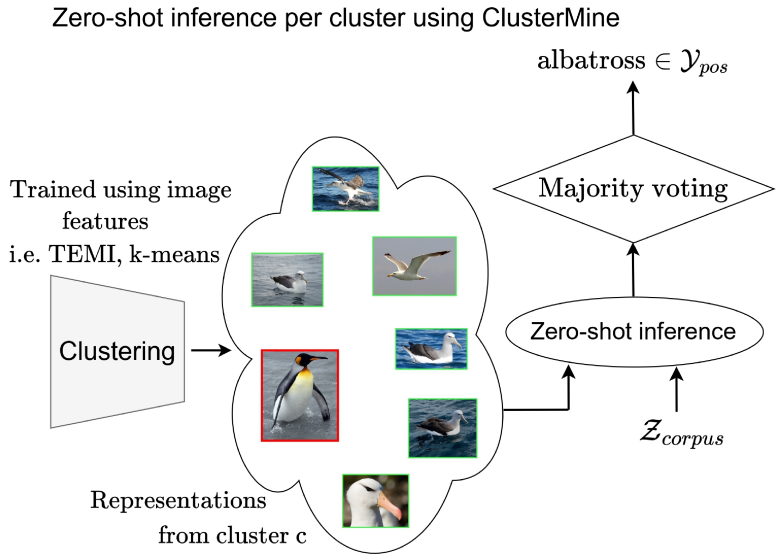


Figure 8.2: **A visual illustration of ClusterMine.**

Cluster-based positive mining (ClusterMine)

The proposed method, cluster-based positive mining (ClusterMine, Fig. 8.2), consists of the following steps:

1. **Visual feature-based clustering:** We perform clustering on the visual encoder of CLIP using C clusters. In practice, we apply TEMI clustering [3] as it has shown significant improvements in clustering accuracy over k-means [35], even at large scales [4]. We use the default parameters ($\beta = 0.6$, 50 heads) as in [3]. In contrast to the clustering downstream task, we are only interested in a rough overestimation of C.

2. **Vision-language inference:** For all samples that fall into the same cluster, we apply zero-shot inference using the text corpus $\mathcal{Y}_{\text{corpus}}$.

3. **Cluster Voting:** Each cluster's label name is then determined by applying majority voting, effectively reducing the false positive classes. The latter enforces visual consistency, as the nearest neighbors in feature space likely share the same label [56, 3]. Crucially, because different clusters can be mapped to the same label name, $|\mathcal{Y}_{\text{pos}}| \leq C$. A heuristic for setting C is the "elbow" method [41] using the saturation of the ratio $|\mathcal{Y}_{\text{pos}}|/C$ or the percentage of $|\mathcal{Y}_{\text{pos}}|$ that appear on multiple clusters as the primary metric, analogous to [41, 1].

ClusterMine has the following advantages over PosMine. It additionally accounts for image-image similarities within the ID features via clustering, similar to a human. By integrating visual consistency into the top-1 image-text concept from $\mathcal{Y}_{\text{corpus}}$: 1) different clusters can be mapped to the same label name $y \in \mathcal{Y}_{\text{pos}}$, and 2) text concepts that do not match the samples' neighborhood are rejected (Fig. 8.2). Thus, $|\mathcal{Y}_{\text{pos}}|$ becomes relatively insensitive as C increases far above $\mathcal{Y}_{\text{real}}$ (Fig. 8.7, right), unlike M in PosMine (Fig. 8.7, center). In practice, an overestimation of the real semantic categories for selecting C is possible even with minimal to no domain knowledge [56, 1]. Experimental

Method	NINCO	IN-O	OpenImage-O	iNat	IN-OOD	Textures43	Avg AUROC/ FPR95
Methods requiring in-distribution label names $\mathcal{Y}_{GT}$ for zero-shot inference							
MaxLogit [22]	83.98/58.03	90.69/39.85	92.52/35.54	92.23/41.80	90.72/44.01	87.77/51.81	89.65/45.17
MD [35, 48]	85.22/64.99	91.37/43.65	91.01/59.68	87.20/90.30	90.72/50.45	94.76/31.14	90.05/56.70
Relative MD	89.25/49.49	92.18/35.50	94.32/29.08	91.58/52.79	89.53/45.21	90.61/43.40	91.25/42.58
MCM [43]	88.78/49.00	91.30/41.20	96.64/16.80	96.62/16.15	89.65/47.36	91.75/40.84	92.46/35.22
PLP [2]	91.80/42.57	93.30/33.05	**97.87/10.84**	98.94/3.86	90.01/47.42	**94.79/26.01**	94.45/27.29
Negative label mining from $\mathcal{Y}_{corpus}$ *given* $\mathcal{Y}_{GT}$							
NegLabel	88.7/50.12	85.29/55.7	94.61/26.08	99.4/**2.46**	82.72/62.62	85.06/56.19	89.3/42.19
NegLabel*	90.26/47.07	90.10/43.45	95.79/22.31	98.64/6.75	88.4/48.34	90.44/43.95	92.27/35.31
CLIPScope	92.69/39.47	88.18/45.15	96.24/17.66	**99.45**/2.57	91.04/43.15	90.84/43.77	93.07/31.96
Ours: Positive label mining from $\mathcal{Y}_{corpus}$							
PosMine	92.56/33.53	93.13/32.3	97.04/15.8	98.83/5.83	91.36/38.89	93.81/32.1	94.46/26.41
ClusterMine	**92.87/30.3**	**93.57 / 29.4**	96.93/15.9	99.0/4.77	**91.53/38.26**	93.45/32.89	**94.56/25.26**

Table 8.1: **Semantic large-scale OOD detection AUROCs (↑)/FPR95(↓) per dataset using CLIP ViT-H dfn5b [16]**. The WordNet [42] corpus is used (nouns and adjectives), and the ID is ImageNet. All reported baselines are reproduced results and do not require training or fine-tuning of CLIP. MD stands for Mahalanobis distance [35]. The best scores are shown in **bold**. The symbol $*$ indicates tuning $|\mathcal{Y}_{\text{neg}}|$ to 40000 for NegLabel [33], which is different from the authors' choice of 10000.

results show a superior label quality for ClusterMine (Fig. 8.5,section D.3). Compared to MCM, ClusterMine leverages a corpus to decide which concepts are positive and negative. Compared to NegLabel [33], ClusterMine extracts the positives and *implicitly* defines the negative concepts without a pre-defined explicit threshold of NegLabel, which is hard to determine a priori.

Combining positive and negative label mining. PosMine and ClusterMine can be easily combined with the negative mining strategy from Section C.1. After computing $\mathcal{Y}_{\text{neg}}$ using Eq. (8.2), the K most dissimilar text representations from $\mathcal{Y}_{\text{pos}}$ are calculated. Under this prism, previous methods [17, 33] can be viewed as special cases of our label mining framework (Fig. 8.1). Interestingly, we show that state-of-the-art large-scale CLIP models do not require pruning $\mathcal{Y}_{\text{neg}}$ when $\mathcal{Y}_{\text{pos}}$ are extracted from the corpus. Unless otherwise specified, PosMine and ClusterMine default to using $\mathcal{Y}_{\text{neg}}$ as shown in Eq. (8.2).

D Experimental evaluation

D.1 Training-free OOD detection baselines

We focus on approaches that do not require training the image or text encoder of CLIP or additional text encoders [61]. To this end, we consider the following CLIP-based OOD detection methods using $\mathcal{Y}_{\text{GT}}$ as baselines: Energy [47], MaxLogit [22], Mahalanobis distance (MD), relative MD [35, 49], MCM [43], and PLP [2]. For instance, MCM defines $S(x)$ as the maximum softmax probability (MSP) of image-text similarities. PLP [2] uses the same pipeline as in MCM, but derives a pseudo-label for each image and trains a linear classifier using pseudo-labels.

We adopt NegLabel [33] as our primary baseline method, which utilizes an external corpus. We highlight that only ClusterMine and PosMine do not assume access to

$\mathcal{Y}_{\text{GT}}$. We use $K = 4 \cdot 10^4$ negative concepts instead of $K = 10^4$ used in [33], which we denote as NegLabel*. Finally, we show a small-scale comparison using ViT-B with recent state-of-the-art methods, such as AdaNeg and SynOOD [70, 36], which use additional information or data.

D.2 Implementation details

Visual OOD datasets. Using ImageNet as ID, we carefully choose six OOD datasets with a sufficiently small degree of semantic overlap based on previous literature. Specifically, we include NINCO [9], IN-O [27], OpenImage-O [59], a subset of plant images from iNaturalist (iNat) [30], IN-OOD [66], and a subset of Textures called Textures43 [59]. IN-OOD and IN-O are constructed from ImageNet-21K, excluding ImageNet samples. NINCO and IN-OOD have a lower degree of semantic overlap, due to their careful manual sample selection. Unless otherwise specified, the mean AUROC and FPR95 are computed from these six out-distributions.

Covariate shifted data. We consider five commonly used covariate-shifted ID in generalization benchmarks, namely ImageNetV2 [47], ImageNet-C [24], ImageNet-A [28], sketches [58], ImageNet-R [23]. ImageNet-R consists of 30000 samples from 200 ImageNet classes, where images contain stylistic and rendition variations, including cartoons, graphics, sketches, and tattoos. ImageNetV2 consists of independently collected samples from Flickr. ImageNet-A is collected using adversarial filtration from supervised classifiers, while ImageNet-C applies 15 pixel-level manipulations, such as adding blurring or weather-like effects. We use a randomly selected subset of 10000 images of ImageNet-C.

Text corpora. We utilize WordNet [42] as a representative large-scale text corpus. We use all nouns and adjectives by default, resulting in $|\mathcal{Y}_{\text{corpus}}| \approx 79 \times 10^3$ concepts. We also adopt the preprocessed ImageNet-21K (IN21K) [50] subset. As a larger-scale corpus, we include Part-of-Speech (POS) Taggings [8], which contains 270×10^3 concepts. We create subsets by considering only nouns (N) or nouns and adjectives (NA). We pre-process the corpus by removing duplicates and using one lemma of the SynSet in Wordnet [42]. Other choices for text pre-processing had similar results.

Vision language models, runtime and memory usage. We use CLIP ViT-H dfn5b [16] weights by default. For the cross-model analysis, we use 12 publicly available weights from [46, 67, 31, 11, 16]. We set τ to be 10 times smaller than the pretraining temperature used in each model. In our experiments, we set $C = 4000$ clusters and $M = 100$ and present their sensitivity in Fig. 8.7. For clustering, we use TEMI [3] with the default parameters. ClusterMine remains training-free w.r.t the CLIP parameters, so image and text features can be precomputed. With the largest scale model ViT-G, clustering on ImageNet takes less than 1.5 hours using TEMI on a single GPU with less than 10 GB of VRAM using the default settings. Since text and image representations are pre-computed, inference for all six OOD detection benchmarks can be achieved in a *maximum* of twelve minutes with less than 16GB of VRAM for a mini-batch size of 4096 for ViT-G with the POS corpus.

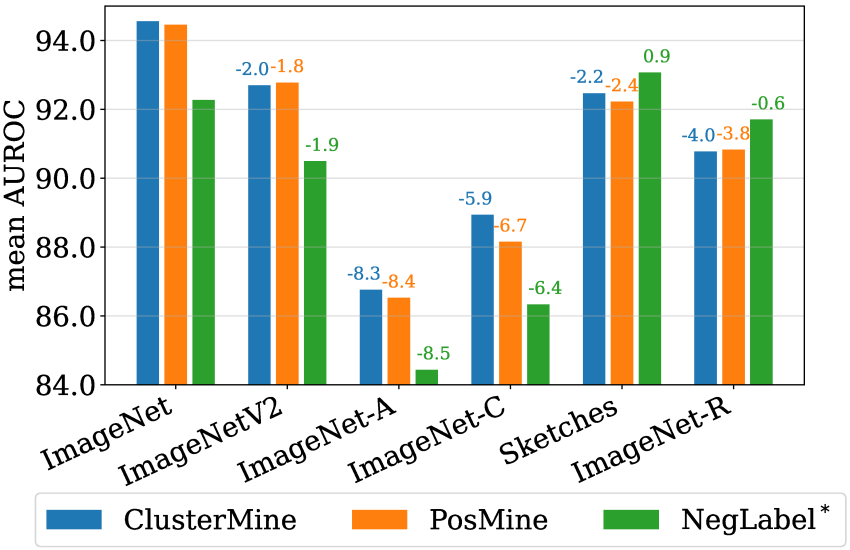


Figure 8.3: **OOD detection robustness to multiple ID shifts (x-axis) compared to ImageNet using CLIP ViT-H dfn5b [16]**. The relative AUROC difference in % of each method compared to its ImageNet score is shown on top of each bar. We report the mean AUROC (↑,%) across six different OOD datasets (y-axis).

D.3 Main experimental results

Main results. ClusterMine is the first method the achieves state-of-the-art OOD detection on most ImageNet benchmarks without $\mathcal{Y}_{\text{GT}}$ (Table 8.1). Interestingly, ClusterMine outperforms in the most curated benchmarks that have the least amount of semantic overlap, namely NINCO, IN-O, and IN-OOD. This reveals the ability of large-scale vision language models to *dynamically* extract positive concepts, given a sufficiently descriptive corpus. This is crucial when in-distribution drifts occur naturally [57].

Scalability. Here, we investigate if positive label mining methods remain competitive across various scales of pretrained networks (Fig. 8.4). We refer to scale as the combination of computing resources, model size, and pretraining data. Apart from model size (i.e., number of parameters), existing models differ in terms of compute time and training data. For this reason, we rank the models based on their OOD detection using ClusterMine in Fig. 8.4. We observe that ClusterMine consistently outperforms NegLabel and MCM across 12 model weights by sizable margins in AUROC. Our results suggest that the ability of CLIP to capture ID-related concepts does not depend on the scale.

Covariate ID shifts and corpus sensitivity. By varying the ID while keeping the same OOD data, we measure the average AUROC degradation, as shown in Fig. 8.3. Both ClusterMine outperform NegLabel on 4 out of 6 IDs. Specifically for the image-level corrupted version of ImageNet (ImageNet-C), we find that ClusterMine is the most robust method for pixel-level perturbations. This suggests that including $\mathcal{Y}_{\text{neg}}$ using Eq. (8.1) is likely the primary driver of OOD detection robustness, similar to proximal works [17, 61, 37, 33, 70].On Sketches and ImageNet-R, NegLabel seems slightly more robust as measured by AUROC when stylistic changes occur in the ID in Fig. 8.3, whereas ClusterMine is superior when comparing FPR95, as shown in Section D.3. While stylistic concepts coexist in the WordNet corpus, we find that the superior AUROC of NegLabel

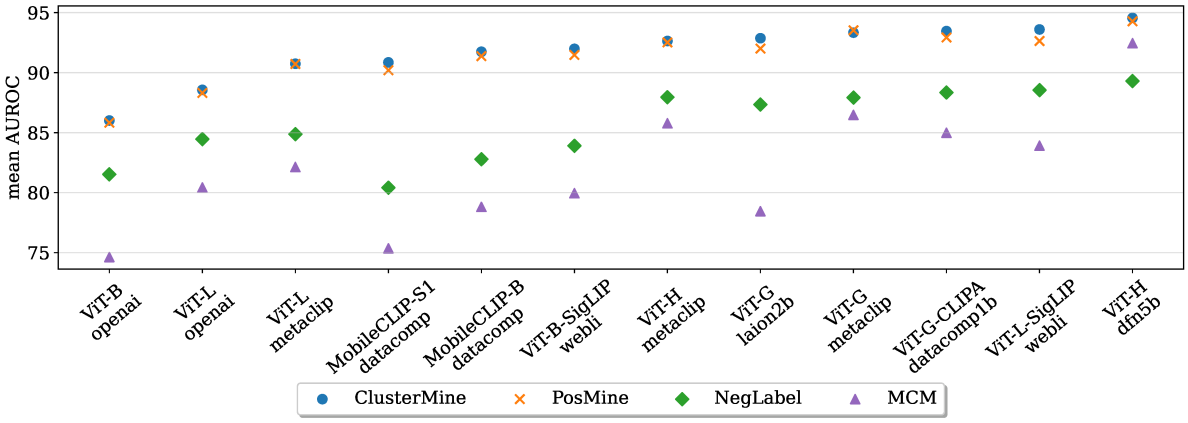


Figure 8.4: **Scalable out-of-distribution detection AUROC (%, y-axis) using various pretrained CLIP weights (x-axis)**. Unlike previous state-of-the-art methods that require $\mathcal{Y}_{\text{GT}}$ (MCM, NegLabel), *ClusterMine and PosMine* extract the in-distribution-related label names from a text corpus. Mean AUROC is computed across six OOD datasets, using ImageNet-1K as ID. We use the WordNet [42] corpus. Pretrained CLIP models are sorted based on their performance with respect to ClusterMine.

ID Dataset	Method	Average AUROC	Average FPR95
ImageNet-Sketches	NegLabel*	**93.06**	40.29
	PosMine	92.22	32.92
	ClusterMine	92.47	**30.94**
ImageNet-R	NegLabel*	**91.70**	41.42
	PosMine	90.82	**35.92**
	ClusterMine	90.78	37.50
ImageNet-A	NegLabel*	84.43	56.32
	PosMine	86.52	44.44
	ClusterMine	**86.76**	**42.57**
ImageNetV2	NegLabel*	90.49	45.38
	PosMine	**92.78**	32.15
	ClusterMine	92.70	**31.87**
ImageNet-C	NegLabel*	86.33	54.87
	PosMine	88.15	52.10
	ClusterMine	**88.94**	**49.21**

Table 8.2: **OOD detection robustness to covariate shifts**. The star symbol indicates tuning $|\mathcal{Y}_{\text{neg}}|$ to 40K for NegLabel.

Corpus	$\lvert\mathcal{Y}_{\text{corpus}}\rvert$	NegLabel*	PosMine			ClusterMine		
		AUROC	AUROC	Overlap	F1 score	AUROC	Overlap	F1 score
WordNet subsets								
IN10K	10K	92.07	95.06	95.6	61.1	**94.54**	87.4	76.3
IN21K	20K	92.38	**95.12**	93.5	57.9	94.66	83.9	71.1
N	67K	92.31	94.39	87.7	52.4	**94.44**	77.2	61.4
NA	79K	92.27	94.46	87.1	51.9	**94.56**	76.0	60.0
POS subsets								
N	231K	91.75	92.79	33.2	19.3	**92.94**	27.8	21.3
N$\cup\mathcal{Y}_{\text{GT}}$	231K	91.26	93.95	78.8	46.6	**94.14**	69.5	53.5
NA	270K	91.78	92.69	33.8	19.7	**92.90**	28.4	21.6
NA$\cup\mathcal{Y}_{\text{GT}}$	270K	91.27	94.01	79.1	46.9	**94.10**	70.0	53.7

Table 8.3: **Average OOD detection AUROC across 6 OOD datasets using various text corpora and CLIP VIT-H dfn5b [16]**. The subsets "N" refer to nouns only, and NA refers to nouns and adjectives. For the POS corpora, we show the impact of manually adding the missing $\mathcal{Y}_{\text{GT}}$ ($\cup\mathcal{Y}_{\text{GT}}$). The class overlap (in %) is defined as $|\mathcal{Y}_{\text{GT}} \cap \mathcal{Y}_{\text{pos}}|/|\mathcal{Y}_{\text{GT}}|$ and the F1 score (in %) is measured between $\mathcal{Y}_{\text{pos}}$ and $\mathcal{Y}_{\text{GT}}$.

on ImageNet-R and Sketches is not attributed to the negative label mining.

As illustrated in section D.3, ClusterMine shows superior performance in all but one corpus. Even when using the complete POS corpus with 270K concepts, we report competitive performance with PLP [2], the best-performing method using $\mathcal{Y}_{\text{GT}}$. Manually adding the missing $\mathcal{Y}_{\text{GT}}$ labels on the POS corpora improves the OOD detection AUROC for ClusterMine , suggesting that domain-specific concepts are beneficial. Although this might be a limitation for novel application domains, it is not necessary to outperform NegLabel in section D.3.

Mined label quality and controlled experiments

Since $|\mathcal{Y}_{\text{pos}}|$ is implicitly determined, we find the F1 score between $\mathcal{Y}_{\text{pos}}$ and $\mathcal{Y}_{\text{GT}}$ a more suitable comparison compared to the percentage of overlapping classes ($|\mathcal{Y}_{\text{GT}} \cap \mathcal{Y}_{\text{pos}}|/|\mathcal{Y}_{\text{GT}}|$). Interestingly, ClusterMine achieves a superior F1 score for all text corpora (section D.3). This is primarily attributed to the cluster voting strategy, leading to *fewer false positives*. In addition, we visualize the semantic alignment of $\mathcal{Y}_{\text{pos}}$ with $\mathcal{Y}_{\text{GT}}$ using two normalized histograms. In Fig. 8.5 left, we measure the top-1 cosine similarity using CLIP text representations to $\mathcal{Y}_{\text{GT}}$, and in Fig. 8.5 right, we measure the shortest path (number of hops) in WordNet. Additionally, intercluster purity remains high ($> 50\%$), while the percentage of mined $|\mathcal{Y}_{\text{pos}}|$ that appear across multiple clusters increases as C increases, confirming that ClusterMine maintains semantic consistency while being robust to the overestimation of C.

In Table 8.4, to assess whether the higher cardinality of $\mathcal{Y}_{\text{pos}}$ alone (compared to $\mathcal{Y}_{\text{GT}}$) is the mere reason for superior performance, we compare i) $|\mathcal{Y}_{\text{pos}}| \approx |\mathcal{Y}_{\text{GT}}|(1000)$ for ImageNet by setting C to $1.2K$ for ClusterMine, and ii) $|\mathcal{Y}_{\text{pos}}| = 1500 = |\mathcal{Y}_{\text{GT}} \cup \mathcal{Y}_{\text{aug}}|$. To mine additional positive labels $\mathcal{Y}_{\text{aug}}$ using $\mathcal{Y}_{\text{GT}}$, we augment $\mathcal{Y}_{\text{GT}}$ with the 500 most similar labels from the corpus using top cosine similarity, and compare with MCM and using Eq. (8.1) as an OOD score. Results are reported in Table 8.4. Intriguingly, increasing

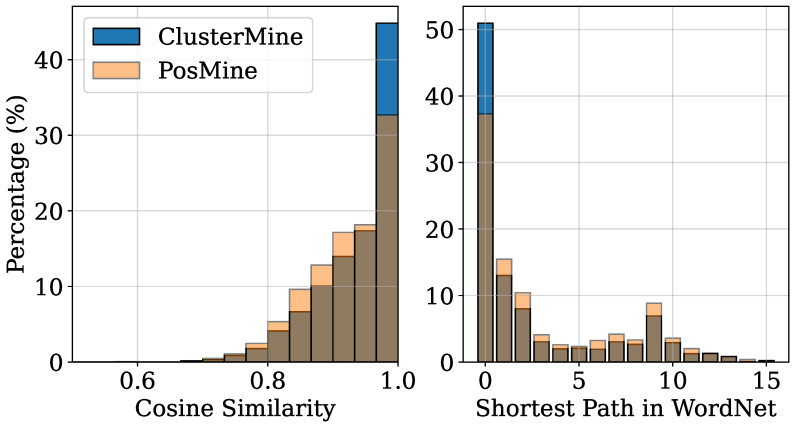


Figure 8.5: **Analysis of mined label name quality.** We calculate top-1 text-text similarity with GT (left), and by finding the shortest path (minimum amount of hops) to GT in WordNet (right).

$\lvert\mathcal{Y}_{\text{pos}}\rvert$:	1000	1500
MCM ($\mathcal{Y}_{\text{GT}} \cup \mathcal{Y}_{\text{aug}}$)	92.46/35.22	92.01/36.17
Eq.1 ($\mathcal{Y}_{\text{GT}} \cup \mathcal{Y}_{\text{aug}}$)	93.08/32.00	92.88/32.81
ClusterMine (C=1.2K, 4K)	**93.84/30.74**	**94.56/25.26**

Table 8.4: **Average AUROC/FPR95 for equal $\lvert\mathcal{Y}_{\text{pos}}\rvert$.** In the first two rows, additional positive label names are mined from $\mathcal{Y}_{\text{GT}}$.

$|\mathcal{Y}_{\text{pos}}|$ alone is **not** improving performance (Table 8.4) and using the $\mathcal{Y}_{\text{GT}}$ positives as anchor does not increase OOD detection performance.

Cluster-based metrics for ClusterMine. To further investigate the quality of the clusters, we compute in Fig. 8.6: (i) the intercluster **purity** (percent of samples withing a cluster that share the majority label), (ii) the intercluster **entropy** w.r.t. mined labels $\mathcal{Y}_{\text{pos}}$, and (iii) how often (percentage) $\mathcal{Y}_{\text{pos}}$ appear across multiple clusters (redundancy ratio), (iv) the ratio of mined $|\mathcal{Y}_{\text{pos}}|/C$. We find that clusters typically exhibit high purity relative to the number of clusters ($\geq 50\%$). This further supports our assumption that feature-space neighbors are label/cluster-consistent. Interestingly, the redundancy ratio increases as C increases, confirming that ClusterMine maintains semantic consistency while being robust to the overestimation of C. This analysis highlights the benefits of choosing ClusterMine over PosMine or NegLabel, where it is challenging to determine their respective hyperparameters in advance.

D.4 Comparison with recent state-of-the-art

While not an apples-to-apples comparison, in Table 8.5, we report results with ViT-B using ClusterMine with AdaNeg [70] (adapts negatives based on incoming OOD samples) and SynOOD [36], which requires synthetic sample generation and training. Still, ClusterMine outperforms existing state-of-the-art methods on near-OOD benchmarks without a) access to $\mathcal{Y}_{\text{GT}}$, b) synthetic sample generation, c) inference-based adaptations based on available OOD data.

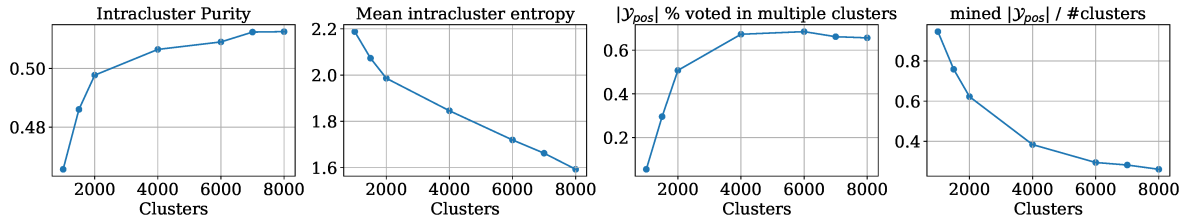


Figure 8.6: **ClustMine cluster analysis for various cluster sizes.** From left to right: intracluster purity measures the percentage of samples within a cluster that have the most frequent (majority) label (*left*), intracluster entropy is computed within samples in the same cluster (*center left*), and we compute the percentage of mined $|\mathcal{Y}_{\text{pos}}|$ that appears across multiple clusters as C increases (*center right*), and finally we compute the ratio of the mined $|\mathcal{Y}_{\text{pos}}|$ related to the chosen number of clusters C (*right*).

Method	Near OOD				Far OOD						Mean	
	SSB-hard		NINCO		iNat		Textures		OpenImage-O		(Near/Far)	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
AdaNeg	74.91	75.11	60.10	78.30	**0.72**	**99.72**	**21.40**	**95.71**	**29.81**	**93.87**	67.5/17.3	76.7/96.4
SynOOD	-	-	-	-	1.57	99.57	22.94	95.29	-	-	71.7/17.1	77.6/96.2
ClusterMine	**68.51**	**76.33**	**54.01**	**84.16**	18.01	96.30	67.69	79.97	30.72	92.55	**61.3**/38.8	**80.3**/89.6

Table 8.5: **ClusterMine comparison using the CLIP ViT-B weights** [46]. We evaluate using the OpenOOD Near- and Far-OOD benchmarks compared to recent state-of-the-art methods. Results for AdaNeg [70] and SynOOD [36] are taken from the reported results.

D.5 Ablation studies and additional OOD datasets

Is negative label mining necessary? Jiang et al. [33] suggest that "the performance of the OOD detector is enhanced by incorporating more negative labels". Their theoretical finding is in contrast with their negative mining strategy. In Fig. 8.7 (left), we show the impact on AUROC from pruning negative labels from the text corpus for ViT-B [46] and the best-performing ViT-H dfn5b [16]. AUROC scores deteriorate from negative label pruning.

Hyperparameter sensitivity. Here, we vary the number of minimum samples per class M for PosMine in Fig. 8.7 (center) and the number of clusters C for ClusterMine (right). As shown in Fig. 8.7, a critical aspect of ClusterMine is its *insensitivity to* C due to the cluster voting step. Other recent methods, such as NegLabel, AdaNeg, and CLIPScope [18] require careful hyperparameter selection, which necessitates an OOD validation set and is challenging to determine for an arbitrary ID and text corpus.

Additional OOD datasets. In section D.5 we report the AUROC on four additional OOD datasets. The Places dataset has the highest semantic overlap with the ID ($\approx 60\%$), while NINCOv2 has a near-zero semantic overlap, as it is a manually picked collection from existing OOD datasets. In contrast to prior works, we use Places as a bad benchmark to showcase how OOD detectors can reject samples that are more likely to be ID. We observe that MCM is the best-performing approach on Places and the worst-performing approach on NINCOv2, respectively. These two benchmarks could be utilized as OOD validation sets in future work.

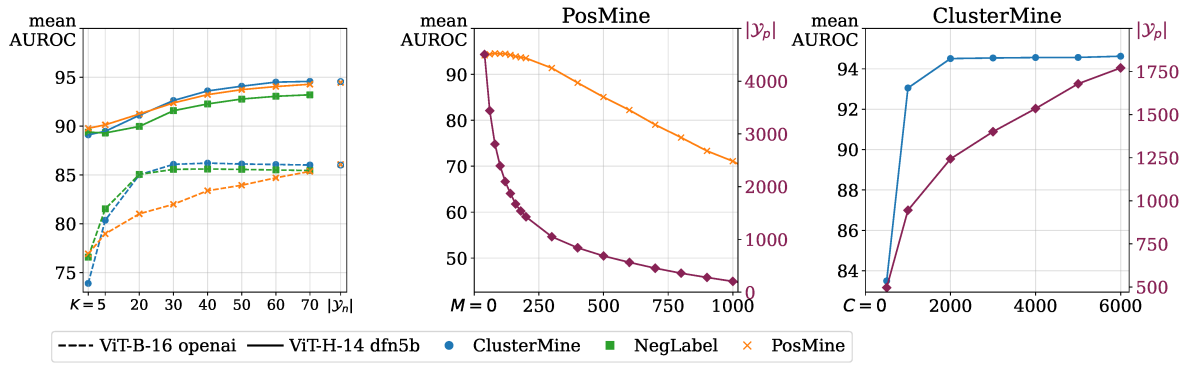


Figure 8.7: **Sensitivity to hyperparameters for negative label mining for all methods (left), PosMine (center), and ClusterMine (right).** We vary the number of negative labels K (in thousands, **left**), minimum number of assigned samples per class M using PosMine (**center**), and number of clusters used in ClusterMine (**right**). The mean AUROC is always reported on the left y-axis. The right y-axes show the cardinality of mined positive label names ($|\mathcal{Y}_{\text{pos}}|$) for different M, C. For $K = |\mathcal{Y}_{\text{neg}}|$ (left plot), all non-overlapping labels from the corpus are used (Eq. (8.2)). Colors and markers per method are shared for all figures. Best viewed in color.

Method	Places	Texture	SUN	SSB	NINCOv2
MCM	**92.32**	90.40	94.37	79.69	92.50
PLP	92.15	**93.08**	94.01	81.42	94.72
NegLabel	90.08	83.13	93.70	82.80	93.07
NegLabel*	90.39	88.29	94.41	85.13	94.53
PosMine	92.08	92.45	**95.51**	85.44	95.58
ClusterMine	91.41	91.80	94.70	**86.04**	**95.86**

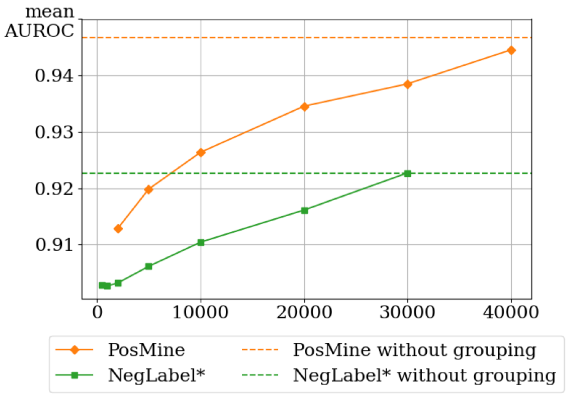
Table 8.6: **Semantic OOD detection detection AUROC on additional OOD datasets.**

Negative label mining on POS corpus and negative grouping strategy. Fig. 8.8b shows that negative label mining is not improving performance even for a large-scale text corpus such as POS. We included all available nouns and adjectives. In Fig. 8.8a, we show that the proposed negative grouping strategy by NegLabel [33] deteriorates the OOD detection AUROC. Thus, we did not include it in our experiments.

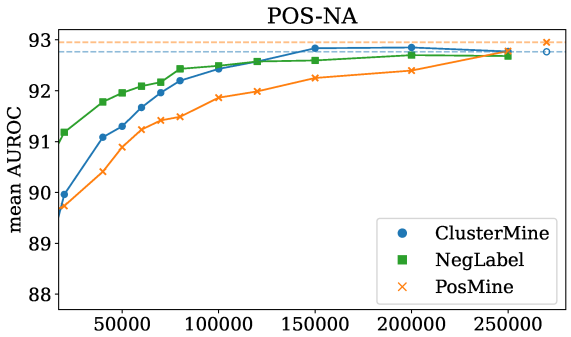
E Future work and conclusion

We believe our work makes a significant step towards *truly unsupervised OOD* using vision-language models by highlighting the importance of dynamically extracting suitable $\mathcal{Y}_{\text{pos}}$. Inarguably, a domain-relevant corpus is easier to satisfy than the exact knowledge of a fixed $\mathcal{Y}_{\text{GT}}$. When no corpus exists, future work could explore the use of a large language model with a human in the loop to generate a domain-specific corpus for other application domains. This will reduce dependence on pre-existing text corpora.

In this work, we presented a general OOD detection label mining framework using



(a) **Ablation study on the negative grouping strategy**. The x-axis represents group size (i.e. number of label names per group).



(b) **The impact of negative label mining on the POS-NA corpus.** The x-axis represents number of mined negative concepts from the corpus.

Figure 8.8: **Comparison of grouping strategy and negative label mining using CLIP VIT-H dfn5b [16]. NegLabel* uses 40K negative mined labels.**

CLIP. We demonstrate that vision language models are able to extract ID-related concepts from a text corpus without relying on $\mathcal{Y}_{\text{GT}}$. We introduced ClusterMine , which is robust to both covariate shifts and variations in the text corpus. ClusterMine achieves state-of-the-art OOD detection scores across a wide range of large-scale OOD benchmarks, all while requiring no fine-tuning in the learned weights of CLIP models.

References

[1] Nikolas Adaloglou, Tim Kaiser, Felix Michels, and Markus Kollmann. Rethinking cluster-conditioned diffusion models. *arXiv preprint arXiv:2403.00570*, 2024.

[2] Nikolas Adaloglou, Felix Michels, Tim Kaiser, and Markus Kollmann. Adapting contrastive language-image pretrained (clip) models for out-of-distribution detection. *arXiv e-prints*, pages arXiv–2303, 2023.

[3] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. *arXiv preprint arXiv:2303.17896*, 2023.

[4] Nikolas Adaloglou, Felix Michels, Kaspar Senft, Diana Petrusheva, and Markus Kollmann. Scaling up deep clustering methods beyond imagenet-1k. *arXiv preprint arXiv:2406.01203*, 2024.

[5] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pages 1454–1471. PMLR, 2023.

[6] James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

[7] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[8] Ruchi Bhatia. Part-of-speech tagging kaggle dataset. CC0: Public Domain Licence.

[9] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.

[10] Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International conference on machine learning*, pages 1392–1402. PMLR, 2020.

[11] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[15] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[16] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[17] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

[18] Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad Khorrami. Clipscope: Enhancing zero-shot ood detection with bayesian scoring. *arXiv preprint arXiv:2405.14737*, 2024.

[19] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2023.

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[22] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 2022.

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[25] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[26] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[27] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[29] Laura Hollink, Aysenur Bilgin, and Jacco Van Ossenbruggen. Is it a fruit, an apple or a granny smith? predicting the basic level in a concept hierarchy. *arXiv preprint arXiv:1910.12619*, 2019.

[30] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[31] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[32] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*, pages 15067–15088. PMLR, 2023.

[33] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[35] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[36] Jinglun Li, Kaixun Jiang, Zhaoyu Chen, Bo Lin, Yao Tang, Weifeng Ge, and Wenqiang Zhang. Synthesizing near-boundary ood samples for out-of-distribution detection. *arXiv preprint arXiv:2507.10225*, 2025.

[37] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024.

[38] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[39] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[40] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[41] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, et al. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 international seminar on application for technology of information and communication*, pages 533–538. IEEE, 2018.

[42] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[43] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022.

[44] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *arXiv preprint arXiv:2306.06048*, 2023.

[45] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[48] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection, 2021.

[49] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[50] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[51] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976.

[52] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

[53] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

[54] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[55] Junjiao Tian, Yen-Change Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Exploring covariate and concept shift for detection and calibration of out-of-distribution data. *arXiv preprint arXiv:2110.15231*, 2021.

[56] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[57] Huy V Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.

[58] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[59] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[60] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021.

[61] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

[62] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021.

[63] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024.

[64] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[65] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023.

[66] William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms. *arXiv preprint arXiv:2310.01755*, 2023.

[67] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[68] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

[69] Qingyang Zhang, Qiuxuan Feng, Joey Tianyi Zhou, Yatao Bian, Qinghua Hu, and Changqing Zhang. The best of both worlds: On the dilemma of out-of-distribution detection. *arXiv preprint arXiv:2410.11576*, 2024.

[70] Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. *arXiv preprint arXiv:2410.20149*, 2024.

[71] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3388–3397, 2023.

9. Guiding a Diffusion Model with Itself using Sliding Windows

Outline

A Introduction 1
B Related work 2
C CFG and WMG are fundamentally different 3
D Sliding window guidance (SWG) 5
E Experimental evaluation and discussion 7
F Conclusion 11
References 12
Appendix 15

Bibliographic Information

Adaloglou, N., Kaiser, T., Iaguidin, D., & Kollmann, M. (2024). Guiding a diffusion model with itself using sliding windows. 36th British Machine Vision Conference, BMVC 2025, November 2025, Sheffield, UK.

Corresponding author: *adaloglo@hhu.de*

Copyright Notice

The latest preprint of this article is available at `https://arxiv.org/abs/2411.10257`, as published in the British Machine Vision Conference (BMVC), 2025. Clarification of the copyright was adjusted according to the guidelines of the publisher.

Acknowledgements

Computing resources were provided by the German AI Service Center WestAI under the project name "westai9566". In particular, the authors would like to express their gratitude to Prof. Dr. Timo Dickscheid and Fritz Niesel for their guidance regarding the computational resources used for the project. The authors also wish to thank Kaspar

Senft, Georgios Zoumpourlis, and Felix Michels for their constructive feedback, which greatly enhanced the quality of this manuscript.

Individual Contributions

The core idea of using sliding windows to guide diffusion sampling was developed by the author's supervisor. The 2D toy model construction and the associated mathematical proofs were conducted exclusively by Markus Kollman and Tim Kaiser. The author focused on the experiments of sliding window guidance (SWG). Additionally, the author conducted a comparative analysis with the latest state-of-the-art sampling methods, proposed combining existing guidance methods with SWG, and led the writing of the experimental section. During this process, the author made crucial practical adjustments for latent space models and diffusion transformers. Under the author's initiative, additional compute was acquired to properly benchmark the largest publicly available diffusion models. Critically, the author performed a major revision after the initial rejection of the manuscript. Upon discussion with his supervisor, the author improved the core method by introducing the mask sliding window guidance (M-SWG) variant. The author repeated and redesigned most of the experiments with the improved method based on the initial feedback from the reviewers. Additionally, the author provided new quantitative results (IS score, saturation metric) that better illustrated the benefits of M-SWG. The manuscript was restructured to emphasize the contribution of M-SWG and remove parts that overlapped with concurrent works.

Abstract

Guidance is a widely used technique for diffusion models to enhance sample quality. Technically, guidance is realised by using an auxiliary model that generalises more broadly than the primary model. Using a 2D toy example, we first show that it is highly beneficial when the auxiliary model exhibits similar but stronger generalisation errors than the primary model. Based on this insight, we introduce *masked sliding window guidance (M-SWG)*, a novel, training-free method. M-SWG upweights long-range spatial dependencies by guiding the primary model with itself by selectively restricting its receptive field. M-SWG requires neither access to model weights from previous iterations, additional training, nor class conditioning. M-SWG achieves a superior Inception score (IS) compared to previous state-of-the-art training-free approaches, without introducing sample oversaturation. In conjunction with existing guidance methods, M-SWG reaches state-of-the-art Frechet DINOv2 distance on ImageNet using EDM2-XXL and DiT-XL. The code is available at `https://github.com/HHU-MMBS/swg_bmvc2025_official`.

A Introduction

Diffusion models (DMs) have emerged as a powerful approach for generative tasks, achieving remarkable success in areas such as image synthesis and text-to-image generation [43, 15, 21, 44, 24, 40, 1]. Despite their success, DMs often fail to generate high-quality samples [3] and require guidance techniques to improve visual fidelity (Fig. 9.1). The currently most popular method, classifier-free guidance (CFG), pushes the sample towards regions with higher posterior class probability [46]. Unlike its predecessor, classifier guidance [11], which relies on training an external classifier on labeled noisy images, CFG combines conditional and unconditional denoisers, which can be trained jointly [16].

In the following, we denote by $\boldsymbol{x}$ a noisy image and by $\boldsymbol{\epsilon}(\boldsymbol{x}, t; c)$ and $\boldsymbol{\epsilon}(\boldsymbol{x}, t)$ the class conditional and unconditional noise predictors at timestep t of the denoising process [11]. CFG combines the two noise predictions during sampling using the linear extrapolation scheme

$$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{x}, t; c) = \boldsymbol{\epsilon}(\boldsymbol{x}, t; c) + w[\boldsymbol{\epsilon}(\boldsymbol{x}, t; c) - \boldsymbol{\epsilon}(\boldsymbol{x}, t)] \ , \quad \text{with guidance weight } w > 0. \tag{9.1}$$

Therefore, CFG can be viewed as a linear extrapolation method [44, 7]. Equivalent extrapolation schemes can be found for all diffusion model formulations, such as target prediction or flow matching [21, 41].

Limitations of CFG. Despite the widespread use of CFG in conditional synthesis [33], it comes with notable limitations. First, it increases the training budget. When trained jointly, the unconditional task can consume up to 20% of the computational cost [16]. Additionally, while CFG reduces class mismatch between samples and conditions of the noise predictor [40], this benefit comes at the expense of sample diversity, as sampling focuses on regions with high class probability [22]. Since guidance is a linear extrapolation scheme, the sampling trajectory can overshoot the desired distribution, leading to oversaturated or oversimplified images [26]. Finally, CFG sampling is restricted to class-conditional generation by design.

Weak model guidance. Recently, a new class of guidance methods has been developed that addresses some of the limitations of CFG. These methods utilize the linear extrapolation scheme of CFG, given by

$$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) + w[\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t)], \tag{9.2}$$

in a more generic way. Here, the subscripts *pos* for positive and *neg* for negative refer to the sign of the noise predictors. The idea behind Eq. (9.2) is to extrapolate into high likelihood regions by designing a negative model $\boldsymbol{\epsilon}_{\text{neg}}$ that has an inferior performance compared to a well-performing DM $\boldsymbol{\epsilon}_{\text{pos}}$. Typically, $\boldsymbol{\epsilon}_{\text{neg}}$ is derived from $\boldsymbol{\epsilon}_{\text{pos}}$ by re-training using fewer parameters, architecture-based heuristic manipulations, or shorter training times [2, 22]. We refer to this class of guidance methods as *weak model guidance (WMG)*. While WMG methods seem promising, they often require training additional models, or careful manual selection of specific layers to impair [2].

In this paper, we first introduce a toy example to show that (i) WMG samples closer to high likelihood regions than CFG and (ii) guidance works best if $\boldsymbol{\epsilon}_{\text{neg}}$ makes similar

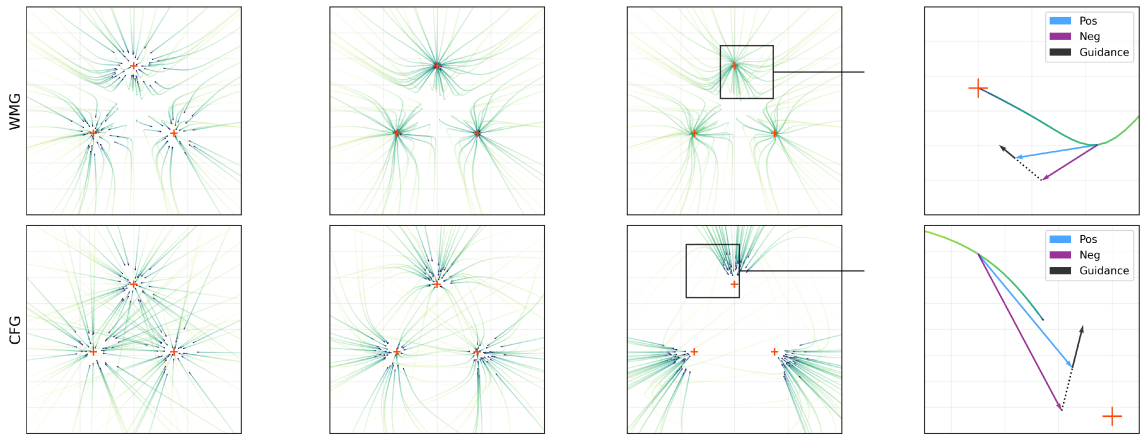
Figure 9.1: **Method overview. Left**: Even state-of-the-art diffusion models can fail to generate globally coherent images without guidance. **Right**: Sliding window guidance (SWG) upweights long-range dependencies and thereby improves global coherence on average.

errors as ϵ_{pos} but stronger. A similar effect can be achieved under increased weight regularization in high-resolution image synthesis. Additionally, we introduce masked sliding window guidance (M-SWG), a novel guidance method designed to upweight long-range dependencies (Fig. 9.1). In contrast to existing guidance techniques [40, 16, 2, 22, 37, 26, 48], M-SWG requires neither training nor class conditioning and can be applied to any DM that can process multiple image resolutions. M-SWG achieves a superior IS score without causing oversaturation, while maintaining competitive Frechet distances. Finally, by combining M-SWG with existing guidance techniques [16, 22], a state-of-the-art Frechet DINOv2 distance (FDD) on ImageNet is achieved using EDM2-XXL and DiT-XL [33, 23].

B Related work

Diffusion sampling can be coarsely divided into CFG-based variations and alternative methods. Vanilla CFG can be improved by adding a noise schedule to the condition [37], introducing monotonically increasing guidance weight schedules [48], stepwise intensity thresholding [40], or applying CFG only at an interval in the intermediate denoising steps [26]. Despite recent advancements, CFG cannot be used with unlabelled datasets or conditional-only trained denoisers by design. While several sampling methods focus on modifying CFG [37, 26, 40, 48, 38], training- and condition-free approaches remain an open quest [39].

Alternative guidance methods can be roughly grouped into three categories: i) *architecture-based impairments*, ii) *image-level manipulations* [18], and iii) inferior capacity models, which we refer to as *weak models*. Architectural impairments typically leverage self-attention maps [2, 17] that are known to capture structure-related information [5, 13, 30, 2]. For instance, a handful of manually picked attention maps can be replaced with an identity matrix [2] or filtered using Gaussian smoothing [17]. Nonetheless, the choice of attention maps depends on the model architecture, which limits the applicability of the method. Finally, architecture-specific impairments can create significant side effects, such as deteriorating the visual structure [17]. Image-level manipulations such as



(a) $w = 0$ (b) $w = w^*$ (c) $w = 3w^*$ (d) Guidance triangle

Figure 9.2: **Inference trajectories for our 2D-toy model. From left to right: (a)** $w = 0$ yields the trajectories of the positive model. **(b)** w^* denotes the guidance weight that leads to best (CFG) or onset of saturating (WMG) performance ($w^* = 1$ for CFG and $w^* = 5$ for WMG). **(c)** For large guidance weights, CFG exhibits large endpoint errors, while WMG keeps small endpoint errors until it becomes unstable. **(d)** Enlarged views for single trajectories. Arrows show target prediction for a point on the trajectory. In this instance, the guidance correction (black arrow) pushes WMG closer to the data point, unlike CFG.

frequency filtering have been attempted with limited success [18]. Hong et al. [18] restrict high-frequency filtering to image patches corresponding to high activation areas, achieved by upsampling self-attention maps. However, similar to architecture-based impairments, it requires cherry-picking while being prone to artifacts [17].

Weak models of inferior capacity compared to the positive model can be constructed by limiting the network size (i.e. number of parameters) or its training time [22]. However, such approaches require additional training from scratch or various weight instances, which are not always available. Concurrent work [4] deteriorates the positive model by fine-tuning with its own generated samples to derive the negative model. Different from prior works [22, 4], we aim to systematically study WMG methods, starting with a toy example. Our analysis reveals that the negative model needs to exhibit *a similar modeling error as the positive model, but stronger.* The latter statement has not been explicitly formulated in prior works [2, 18, 22].

C CFG and WMG are fundamentally different

In the following, we introduce a two-dimensional toy example (Fig. 9.2) to visualize the conceptual differences between CFG and WMG. The data distribution of our toy example consists of three data points $\mathcal{D} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3\}$, where each data point is assigned a different class label $c \in \{1, 2, 3\}$. We can generate trajectories $\boldsymbol{x}(t)$ by numerically solving the ordinary differential equation (ODE) $d\boldsymbol{x}(t) = \boldsymbol{\epsilon}^*(\boldsymbol{x}(t), \sigma(t))d\sigma(t)$, with $\boldsymbol{\epsilon}^*$ the optimal noise predictor in the Bayesian sense, and $\sigma(t) \in [0, \sigma_{max}]$ the current noise level.

Note that $d\sigma(t) < 0$ along a denoising trajectory. It can be shown that for a finite set

of N data points, the optimal noise predictor takes the explicit form [21]

$$\boldsymbol{\epsilon}^*(\boldsymbol{x}, \sigma(t)) = \frac{1}{\sigma(t)} \sum_{i=1}^{N} (\boldsymbol{x} - \boldsymbol{y}_i) p(\boldsymbol{y}_i|\boldsymbol{x}, t), \quad p(\boldsymbol{y}_i|\boldsymbol{x}, t) = \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{y}_i, \sigma(t)^2)}{\sum_{i=1}^{N} \mathcal{N}(\boldsymbol{x}|\boldsymbol{y}_i, \sigma(t)^2)}, \tag{9.3}$$

where $\mathcal{N}(.|.)$ denotes an isotropic normal distribution. That is, the optimal noise predictor is the normalised residual between $\boldsymbol{x}$ and the posterior-mean $\sum_{i=1}^{N} \boldsymbol{y}_i p(\boldsymbol{y}_i|\boldsymbol{x}, t)$ as an estimator for a data point. The optimal noise predictor generates trajectories with endpoints arbitrarily close to one of the data points if the ODE is initialized with $\boldsymbol{x}_{init} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \sigma_{max}^2)$ and σ_{max} is sufficiently large.

Assume we are given an error-prone noise predictor, $\boldsymbol{\epsilon}_{err}$, whose endpoints are distributed as a superposition of three normal distributions, all with variance δ^2, and each normal distribution centered around a different data point (Fig. 9.2a). For this special case it can be shown that $\boldsymbol{\epsilon}_{err}(\boldsymbol{x}, \sigma(t)) = \sigma(t)/\tilde{\sigma}(t)\boldsymbol{\epsilon}^*(\boldsymbol{x}, \tilde{\sigma}(t))$, with $\tilde{\sigma}(t)^2 = \sigma(t)^2 + \delta^2$. The question is now, to what extent CFG and WMG can help to shift the trajectory endpoints of the error-prone predictor closer to the data distribution? To answer this question, we introduce the error-prone predictors $\boldsymbol{\epsilon}_{\text{pos}}$, $\boldsymbol{\epsilon}_{\text{neg}}$, and adjust the guidance weight w of Eq. (9.2) such that the average endpoint error is minimized. Using $\boldsymbol{\epsilon}(\boldsymbol{x}, t) := \boldsymbol{\epsilon}(\boldsymbol{x}(t), \sigma(t))$ for brevity, we recover the CFG setting by introducing the class-conditional noise predictor $\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{err}(\boldsymbol{x}, t; c)$ and the unconditional noise predictor $\boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{err}(\boldsymbol{x}, t)$. Likewise we recover the WMG setting by $\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{err}(\boldsymbol{x}, t)$ and $\boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}'_{err}(\boldsymbol{x}, t)$, where $\boldsymbol{\epsilon}'_{err}$ differs from $\boldsymbol{\epsilon}_{err}$ by a larger endpoint variance $\delta' > \delta$. The denoising trajectories shown in Fig. 9.2 follow from numerically solving the ODE for $\boldsymbol{x}(t)$, using Eq. (9.2) instead of the optimal noise predictor.

CFG. As we use just one datapoint per class, the class conditional noise predictor is given by $\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) = \sigma(t)^{-1}(\boldsymbol{x} - \boldsymbol{y}_{i_c})$, with i_c the data point index belonging to class c, resulting in the special case of straight lines as denoising trajectories (Fig. 9.2a, CFG). For small guidance weights, the trajectory endpoints match slightly better to the data distribution (Fig. 9.2b, CFG), but for larger guidance weights, they move away from the data distribution (Fig. 9.2c, CFG). This behavior is a consequence of the fact that $\boldsymbol{\epsilon}_{\text{pos}}$ drives the trajectory towards the data point with the corresponding class label, whereas $\boldsymbol{\epsilon}_{\text{neg}}$ denoises in the direction of a weighted superposition of all data points. As a result, CFG tends to shift the trajectories away from the center of mass of the data distribution (Fig. 9.2d, CFG), favoring classifiable endpoints over a fit to the data distribution. This behavior has been described in previous works [16] and scales to higher dimensions, e.g. images, where a better distribution overlap between generated samples and training samples in feature space is only observed for sufficiently small guidance weights.

WMG. In contrast to CFG, increasing the guidance weight for WMG drives trajectory endpoints closer to the data distribution (Fig. 9.2a-c, WMG), up to the point where trajectories become unstable. The crucial condition for WMG to work well is that, in each step, the prediction of $\boldsymbol{\epsilon}_{\text{pos}}$ is closer to $\boldsymbol{\epsilon}^*$ than the prediction of $\boldsymbol{\epsilon}_{\text{neg}}$, causing the extrapolation to correct the prediction of $\boldsymbol{\epsilon}_{\text{pos}}$ towards the optimal prediction of $\boldsymbol{\epsilon}^*$ (Fig. 9.2d, WMG).

$$w^*(\boldsymbol{x}, t) = \frac{\|\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}^*(\boldsymbol{x}, t)\|}{\|\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t)\|} \tag{9.4}$$

For the special case that $\boldsymbol{\epsilon}_{\text{pos}}$ and $\boldsymbol{\epsilon}_{\text{neg}}$ exhibit the same error up to a multiplicative factor, it is possible for WMG to closely recover the prediction of the optimal denoiser $\boldsymbol{\epsilon}^*$, under the optimal guidance weight $w^*(\boldsymbol{x}, t)$ (Eq. (9.4)). The optimal weight is obviously of little practical use as it requires knowledge about the optimal noise predictor $\boldsymbol{\epsilon}^*$. However, it illustrates that the practical choice of a constant guidance weight is, in general, an *ad hoc* simplification. Surprisingly, we observe excellent, albeit imperfect, matching between endpoints and datapoints for constant but sufficiently large weights (Fig. 9.2).

D Sliding window guidance (SWG)

Assuming that a significant source of the predictor's error can be attributed to long-range dependencies, we introduce a new guidance method that aims to correct such errors. The idea is to generate $\boldsymbol{\epsilon}_{\text{neg}}$ by restricting the spatial input size ($H \times W$) of $\boldsymbol{\epsilon}_{\text{pos}}$ to a smaller size $k \times l$, with $k < H$ and $l < W$, which induces a defined cut-off for long-range dependencies. We use sliding windows to generate N crops of size $k \times l$, using a fixed stride s per dimension. These crops are independently processed (without rescaling) by $\boldsymbol{\epsilon}_{\text{pos}}$ as illustrated in Fig. 9.3. The N predictions of $\boldsymbol{\epsilon}_{\text{pos}}$ are superimposed in the same order and at the same positions, which results in an $H \times W$ output for $\boldsymbol{\epsilon}_{\text{neg}}$. Overlapping pixels are averaged. We emphasize that using sliding windows is a common technique in high-resolution 2D/3D image processing [8, 20, 42, 36, 6].

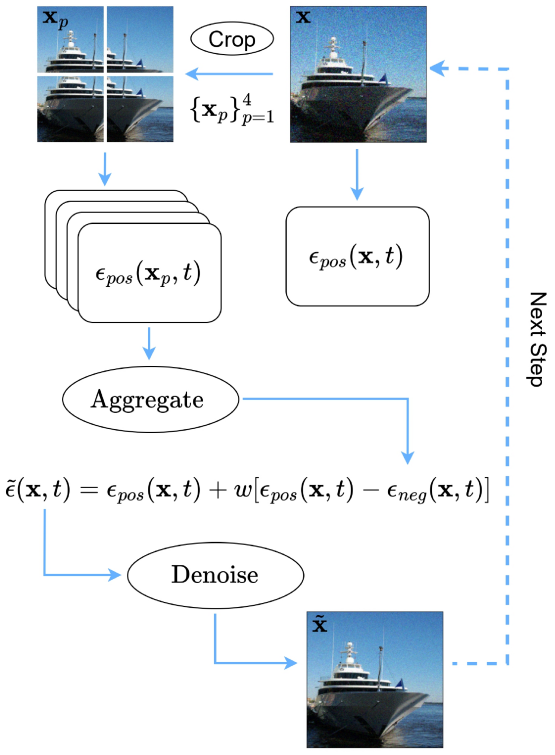


Figure 9.3: **An illustration of sliding window guidance (SWG)**. We use $\boldsymbol{\epsilon}_{\text{pos}}$ to independently process $N = 4$ overlapping image crops. $\boldsymbol{\epsilon}_{neg}$ is generated by superimposing the processed crops and averaging the overlapping pixels.

Hyperparameters and limitations. One limitation of SWG is that $\boldsymbol{\epsilon}_{\text{pos}}$ must process inputs of varying image resolution. Nonetheless, current architectures, such as Unets [15] or (diffusion) image transformers [33], generally satisfy this requirement. SWG can be implemented in a few lines of code and requires no training-related modifications or class conditioning. Unless otherwise specified, we assume $H = W$ and use crop size $k = l \approx 5/8 \times H$ and $N = 4$, which translates into $k = 40, s = 24$ at 64×64 resolution. We denote the overlap ratio per dimension as $r = 1 - \frac{s}{k} = 0.4$.

Model	Guidance	FID	FDD
EDM2 S-512	✗	2.91	112
	$\text{RT}_{.8/1.2}$	**1.79**	46.5
	$\text{SWG}_{.1/.5}$	2.59	56.1
	$\text{M-SWG}_{.2/1.1}$	2.55	53.4
	$\text{RT}_{.8/1.2} + \text{M-SWG}_{0/.25}$	1.79	**40.9**
EDM2 XXL-512	✗	2.29	49.7
	$\text{RT}_{.1/1.3}$	**2.26**	34.6
	$\text{SWG}_{.05/.2}$	2.61	37.7
	$\text{M-SWG}_{.05/.5}$	2.46	36.8
	$\text{RT}_{.1/1.05} + \text{M-SWG}_{0/.1}$	2.26	**32.9**
DiT XL-256	✗	9.83	213
	$\text{SWG}_{.25/.5}$	4.00	80.6
	$\text{M-SWG}_{.5/1.5}$	**3.30**	**70.3**
Unconditional generation			
EDM2 S-512	✗	13.7	250
	$\text{RT}_{1.6/2}$	**4.89**	107.8
	$\text{SWG}_{.3/.5}$	10.2	167.4
	$\text{M-SWG}_{.5/.9}$	9.32	149.4
	$\text{RT}_{1.6/1.35} + \text{M-SWG}_{0/.15}$	4.89	**101.2**

Table 9.1: Training-free diffusion guidance sampling methods comparison.

Model	Guidance	FDD
EDM2 S-512	$\text{RCT}_{1.4}$	42.1
	$\text{RCT}_{1} + \text{SWG}_{.2}$	39.2
	$\text{RCT}_{1.2} + \text{M-SWG}_{.14}$	**39.1**
	$\text{CFG}^{\dagger}_{.8}$	52.9
	$\text{CFG}^{\dagger}_{.3} + \text{M-SWG}_{.3}$	**51.9**
EDM2 XXL-512	RCT_{1}	30.3
	$\text{RCT}_{.9} + \text{M-SWG}_{.05}$	**29.8**
	$\text{CFG}^{\dagger}_{.4}$	36.7
	$\text{CFG}^{\dagger}_{.2} + \text{M-SWG}_{.05}$	**36.2**
DiT XL-256	$\text{CFG}_{2.05}$	54.3
	$\text{CFG}_{.7} + \text{SWG}_{.1}$	50.6
	$\text{CFG}_{.6} + \text{M-SWG}_{.5}$	**49.1**
Unconditional generation		
EDM2 S-512	$\text{RCT}_{1.8}$	97.9
	$\text{RCT}_{1.3} + \text{M-SWG}_{.15}$	**91.8**

Table 9.2: M-SWG enhances existing state-of-the-art methods. The dagger † indicates that a smaller capacity model is used.

Tables 9.1-9.2: We report FDD (↓) and FID (↓) on ImageNet for several benchmarks. The subscripts indicate the chosen guidance scale that yields the optimal guidance weights for FID/FDD based on a hyperparameter search.

SWG in latent DMs. We keep the same design choices for latent DMs [35] that operate in $\frac{H}{8} \times \frac{H}{8}$ feature dimensions, where 8 is the upscaling/downscaling factor. More specifically, SWG is performed in each denoising step in the low-dimensional latent space (i.e. 64×64), and the final output is passed to the decoder. The decoder handles the upscaling to the image dimensions, i.e., from 64×64 to 512×512. The rationale behind SWG operating in lower feature dimensions is that latent representations, i.e., of a VAE, preserve to a significant extent the spatial structure of the image [25, 12]. To make SWG compatible with DiT, we add positional encoding interpolation to process the crops. The choice of $k \approx \frac{5}{8}H$ is determined by practical considerations of Unets. Specifically, k must satisfy $\frac{k}{2^n} \in \mathbb{N}$, where n is the number of downsampling steps of the Unet, and the diffusion model $\boldsymbol{\epsilon}_{\text{pos}}$ must be able to process multiple image resolutions.

Computational overhead. The computational overhead of SWG compared to CFG is determined by the overlap ratio r and the architecture. For EDM-S and $r = 0.4$, we measured an overhead of less than 30% compared to CFG based on a naive implementation of SWG.

Masked SWG (M-SWG). We experimentally observe that applying guidance only to overlapping regions of the sliding windows shows better performance than applying guidance to the whole image. We therefore introduce a binary mask $\mathbf{M} \in \{0,1\}^{H \times W}$ to filter for overlapping regions when applying guidance:

$$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) + w\mathbf{M} \odot [\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}_{\text{neg}}(\boldsymbol{x}, t)]. \tag{9.5}$$

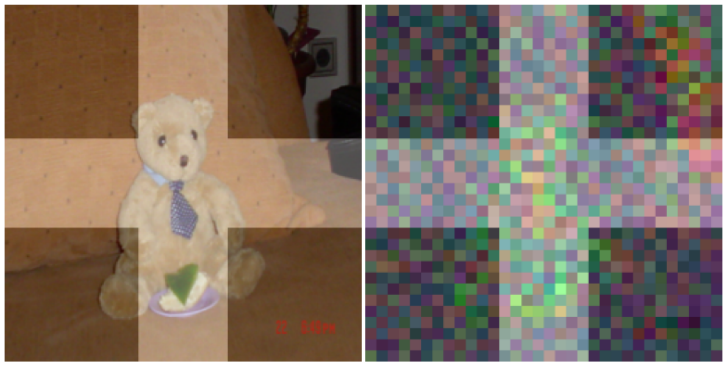


Figure 9.4: **The white overlay (cross) illustrates the pixels that are affected by M-SWG in each timestep in image space (left) and latent space (right), respectively, based on Eq. (9.5).** The overlaid pixel locations corresponds to overlapping regions based on kernel size 40 and stride 24 for an input of size 64×64.

Combining SWG with existing methods. Since SWG and M-SWG are conceptually different from existing techniques, we investigate whether they can enhance existing techniques as a linear combination of their stepwise outcomes [28]. Mathematically,

$$\tilde{\boldsymbol{\epsilon}}(\boldsymbol{x}, t) = \boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) + \sum_i w_i [\boldsymbol{\epsilon}_{\text{pos}}(\boldsymbol{x}, t) - \boldsymbol{\epsilon}^{(i)}_{\text{neg}}(\boldsymbol{x}, t)], \tag{9.6}$$

where i indicates different guidance methods. Quantitative improvements would suggest a complementary relationship between methods. For instance, combining CFG with SWG would increase noise-condition alignment whilst conforming to long-range dependencies.

E Experimental evaluation and discussion

Diffusion guidance baselines. We divide the evaluation into training-free methods (Table 9.1) and state-of-the-art comparisons (Table 9.2), where M-SWG is used in conjunction with existing methods (Eq. (9.6)). Regarding training-free methods, we consider reduced training (RT) as a state-of-the-art baseline [22], along with sampling without guidance ($w = 0$). Nonetheless, publicly available DMs do *not* always offer earlier weight instances such as DiT-XL [33] and StableDiffusion [35], and thus RT cannot be applied. When combining M-SWG with other non-training-free state-of-the-art methods (Eq. (9.6)), we adopt CFG and the recent approach of reducing training time and learning capacity (RCT), also known as *"autoguidance"* [22]. We emphasize that CFG requires

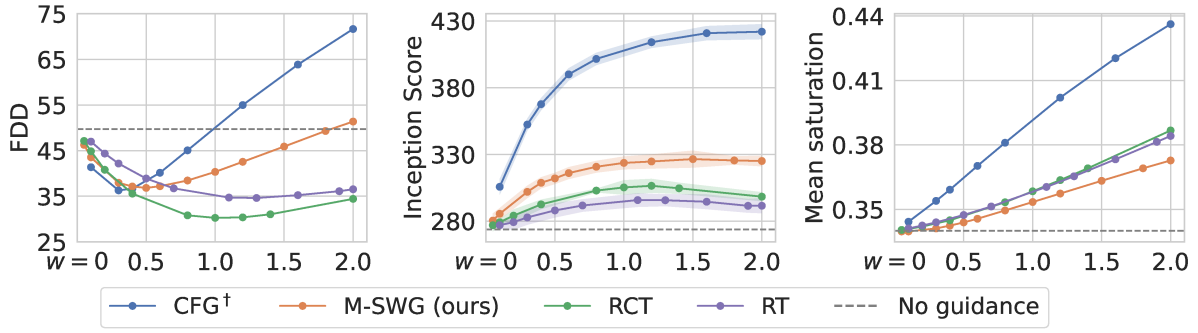


Figure 9.5: **Visualizing the achieved tradeoff by varying the guidance scale w (x-axis) for FDD (↓, left), Inception score (IS ↑, center), and saturation (↓, right) using EDM2-XXL**. M-SWG achieves superior perceptual quality as measured by the IS compared to RT and RCT, while not exhibiting undesired oversaturation as CFG†, on average.

jointly or independently training a conditional and an unconditional denoiser, and it is thus not training-free.

Implementation details. The experiments were conducted on 4 Nvidia A100 GPUs with 40GB of VRAM. Following recent works [45, 23, 19, 26, 22, 9], we produce 50K samples for the evaluations and measure the Frechet distance [14] using features from InceptionV3 and DINOv2 ViT-L [31], denoted by FID and FDD, respectively. We use FDD as our primary metric since it has shown better alignment with human judgment [32, 29, 27, 45]. We also report the inception score (IS) as a measure of visual quality. We apply SWG and M-SWG to state-of-the-art models, such as DiT-XL [33], and EDMv2 [23].

Regarding EDM2, we set the EMA-like hyperparameter to 0.1 to make our analysis more generally applicable. Following Karras et al. [22], we use $T/16$ for EDM2-S and $T/3.5$ for EDM2-XXL for the negative model when applying RT and RCT, where T is the training time of the positive model. When applying CFG to EDM2 models, we use a smaller capacity unconditional model (EDM2-XS), which we denote as CFG†.

M-SWG versus training-free methods. In Table 9.1, we observe significant gains compared to the unguided generation using SWG and M-SWG. M-SWG provides a competitive alternative when previous weight instances, such as with DiT-XL, are unavailable. We believe that the inferior Fréchet distances from SWG/M-SWG are attributed to background simplification, which strongly affects the statistics of the image distribution. This is a shared limitation with CFG [16, 10] and can be further investigated in future work.

Crucially, by combining RT with M-SWG, we achieve a new state-of-the-art FDD for training-free image synthesis using EDM2-XXL (32.9), which also outperforms CFG (36.7). The gains in FDD do not translate to FID. We hypothesize that the less-trained negative model (RT) makes more errors related to local image details, complementing SWG, and it is best captured in the DINOv2 feature space. Finally, no training-free approach can significantly outperform the unguided EDM2-XXL FID score. This is linked to existing concerns that FID does not reflect gradual improvements [32, 27, 45, 19, 1],

N, k, r	FDD	FID
✗	112.2	2.92
4, 32, 0	$55.7_{0.3}$	$2.72_{0.1}$
4, 40, 0.4	$56.1_{0.5}$	$2.59_{0.1}$
4, 48, 0.6	$60.7_{0.5}$	$\mathbf{2.56}_{0.1}$
9, 32, 0.5	$\mathbf{55.4}_{0.3}$	$2.69_{0.1}$
9, 24, 0.63	$58.1_{0.2}$	$2.89_{0.1}$
16,16,0	$64.0_{0.2}$	$3.15_{0.1}$

Table 9.3: Varying kernel size k and number of crops N.

Guidance	Interval	FDD
✗	-	112.2
$\text{RT}_{1.2}$	-	46.5
+ $\text{interval}_{1.8}$	10-20	**43.8**
$\text{SWG}_{0.5}$	-	56.1
+ $\text{interval}_{1.2}$	15-20	51.3
$\text{M-SWG}_{1.1}$	-	53.4
+ $\text{interval}_{1.7}$	13-23	51.7

Table 9.4: Applying guidance only at an interval (32 steps).

	FDD	FID
CFG† [16]	$52.9_{0.8}$	$2.44_{0.1}$
WMG methods		
a) Reduced training	$46.5_{1.2}$	$1.79_{0.8}$
b) Reduced capacity	$70.0_{1.4}$	$2.20_{0.7}$
c) RCT (a + b)	$\mathbf{42.1}_{\mathbf{1.4}}$	$\mathbf{1.67}_{0.8}$
d) WD fine-tuning	$52.1_{0.6}$	$2.46_{0.1}$
e) WD + (a)	$43.9_{0.9}$	$2.25_{0.2}$

Table 9.5: A comparison of WMG strategies, including the novel weight decay (WD) variants.

Tables 9.3-9.5: Additional experiments on ImageNet-512 using EDM2-S. The subscripts indicate the chosen guidance scales based on a hyperparameter search.

as well as that the EDM2-XXL makes fewer errors that can be corrected with guidance techniques.

Similar to EDM2, applying M-SWG to DiT is highly beneficial in terms of FID/FDD. To the best of our knowledge, early training weight instances are not publicly available for DiT to benchmark against RT or other WMG variants. The poor performance of the unguided DiT-XL is mainly attributed to the joint training on both conditional and unconditional generation, which may reduce network capabilities and lead to skewed sampling distributions [22].

State-of-the-art comparison. Table 9.2 shows consistent improvements when combining M-SWG with other state-of-the-art methods. A large relative FDD gain of 9.6% is found when combining M-SWG with CFG using DiT-XL. Compared to using M-SWG in conjunction with RT versus RCT, the relative gain decreases from 4.9% to 1.6%. This suggests that the model of reduced learning capacity partially accounts for the correction in long-range dependencies. Finally, M-SWG shows significant improvements on the unconditional benchmarks. Future work can explore SWG and M-SWG for text-conditioned DMs.

Quality-diversity tradeoffs. To investigate the achieved tradeoff of M-SWG, we plot w against FDD, IS score, and saturation in Fig. 9.5. Saturation is quantified as the mean channel intensity in HSV color space (S channel), as in [38]. While RT and RCT are slightly superior in terms of FDD, M-SWG achieves a better image quality (IS) across guidance scales. Intriguingly, we find that M-SWG does not suffer from oversaturation, unlike the CFG and RCT. A low saturation is preferred because it is closer to the statistics of real images (0.32 on the ImageNet test set). Moreover, the root mean square contrast of the samples (the standard deviation of intensities after greyscale conversion [34, 38]) demonstrates an even more pronounced disparity between M-SWG and other methods. Overall, the smaller FID improvements of M-SWG (e.g. 2.55 versus 2.91 for EDM2-S in Table 9.1), while achieving a higher IS score than RC, RCT suggests that the merits of M-SWG are more in improving semantic coherence than diversity of images.

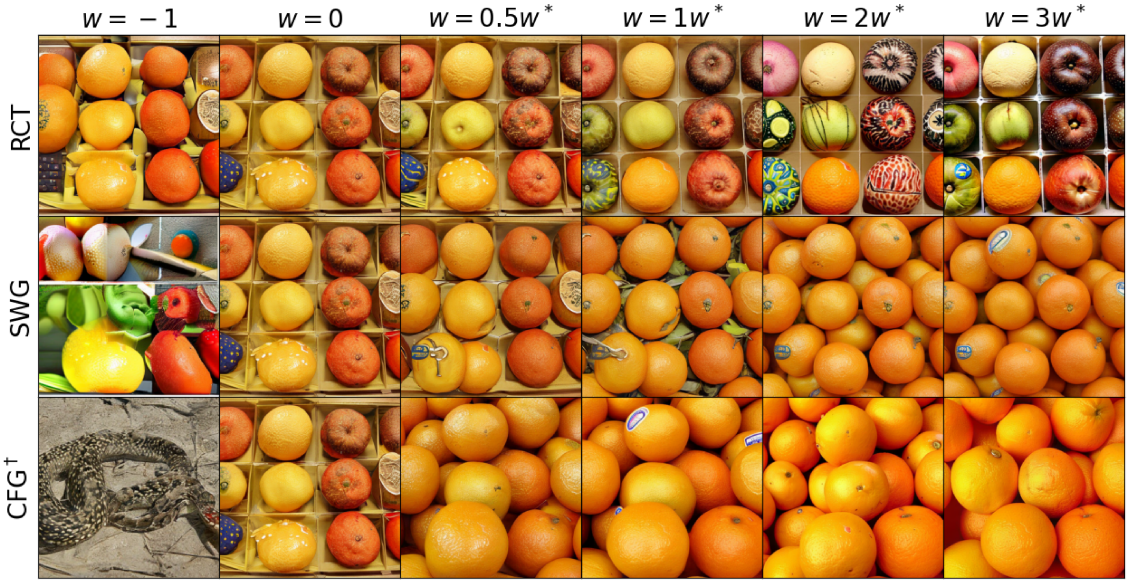


Figure 9.6: **Samples generated by different guidance methods (rows) with different guidance weights (columns) but the same initial noise and condition (orange class) using EDM2-XXL.** Guidance weights w are scaled by w^*, the individual weight for each method corresponding to the lowest FDD. The first column shows samples using $\boldsymbol{\epsilon}_{neg}$, which corresponds to setting $w = -1$ in Eq. (9.2).

Additional experiments. Table 9.3 shows that several combinations of kernel size and N crops can attain similar results on both FID and FDD. Intriguingly, while FID prefers higher overlap ratios r (for $N = 4$), FDD shows that the overlap between crops is not required to obtain improvements over the unguided model. This suggests that (i) even without overlapping crops, SWG introduces no severe artifacts w.r.t. generative metrics, and (ii) minimal to no tuning is required as opposed to heuristic model-specific impairments like SAG or PAG [18, 2]. Moreover, in Table 9.4, we demonstrate that applying guidance only at an intermediate interval further enhances SWG and M-SWG.

Same error but stronger in high dimensions. Motivated by the toy model, we verify its implication in higher dimensions by constructing $\boldsymbol{\epsilon}_{\text{neg}}$ with the same network architecture as $\boldsymbol{\epsilon}_{\text{pos}}$, but with significantly increased weight regularization. This can be implemented by either fine-tuning $\boldsymbol{\epsilon}_{\text{pos}}$, denoted as *Weight decay (WD) fine-tuning*, or by re-training it with very little compute time, which we call *WD with reduced training*. WD increases the model's bias in the context of the bias-variance tradeoff [47]. Naive training-free approaches such as dropout or quantization to construct $\boldsymbol{\epsilon}_{\text{neg}}$ have been unsuccessful [22]. Experimental results are presented in Table 9.5. Interestingly, WD with reduced training performs comparably with the current state-of-the-art RCT, demonstrating that it is another principled way of constructing $\boldsymbol{\epsilon}_{\text{neg}}$.

Qualitative results. To visualize the differences between the guidance methods, we examine ImageNet-512 samples generated from the same initial noise in Fig. 9.6. All three methods, RCT, SWG, and CFG†, tend to exhibit improved perceptual quality for moderate guidance weights but show reduced variation (SWG, CFG†) or artifacts (RCT)

for higher guidance weights. To understand these effects, it is helpful to examine the unguided samples produced by $\boldsymbol{\epsilon}_{\text{neg}}$, which corresponds to setting $w = -1$ in Eq. (9.2). For example, comparing the first two columns of Fig. 9.6, we see that the image composition (e.g. box of fruits) for $\boldsymbol{\epsilon}_{\text{pos}}$ ($w = 0$) is only maintained by $\boldsymbol{\epsilon}_{\text{neg}}$ ($w = -1$) for RCT. Consequently, with increasing w, RCT meaningfully conserves the image composition, SWG does so slightly, and CFG† not at all.

F Conclusion

In this work, we identified the general concept of "same error but stronger" for designing diffusion guidance methods. We validated this concept using a toy example and experimentally in high dimensions using a negative model with increased weight regularization. Additionally, we presented a novel training-free and condition-free guidance method (M-SWG) that enhances global coherence and consistency of objects. M-SWG can be used as a standalone method or combined with state-of-the-art sampling methods, such as CFG, to yield visually appealing and state-of-the-art results in image synthesis.

References

[1] Nikolas Adaloglou, Tim Kaiser, Felix Michels, and Markus Kollmann. Rethinking cluster-conditioned diffusion models for label-free image synthesis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3603–3613. IEEE, 2025.

[2] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. *arXiv preprint arXiv:2403.17377*, 2024.

[3] Md Manjurul Ahsan, Shivakumar Raman, Yingtao Liu, and Zahed Siddique. A comprehensive survey on diffusion models and their applications. 2024.

[4] Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data. *arXiv preprint arXiv:2408.16333*, 2024.

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[6] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

[7] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

[8] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[9] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them, 2020.

[10] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[12] Sander Dieleman. Generative modelling in latent space, 2025.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[17] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *arXiv preprint arXiv:2408.00760*, 2024.

[18] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023.

[19] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

[20] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*, 2023.

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022.

[22] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself, 2024.

[23] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2024.

[24] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023.

[25] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509*, 2025.

[26] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models, 2024.

[27] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance, 2023.

[28] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[29] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for gan evaluation. In *International Conference on Learning Representations*, 2021.

[30] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024.

[31] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[32] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation, 2022.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[34] Eli Peli. Contrast in complex images. *Journal of the Optical Society of America A*, 7(10):2032–2040, 1990.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[36] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024.

[37] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *ICLR*, 2024.

[38] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*.

[39] Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv preprint arXiv:2407.02687*, 2024.

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[41] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022.

[42] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024.

[43] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

[45] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, 2023.

[46] Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025.

[47] Ulrike Von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pages 651–706. Elsevier, 2011.

[48] Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024.

Appendix

App. I Additional implementation details and experiments

Constrast. In Fig. 9.7, similar to saturation, we show that the discrepancy between M-SWG and other state-of-the-art methods is significant. The combination of competitive IS and FDD, with significantly less oversaturation and almost no overcontrast on average, suggests that samples remain relatively closer to the training data distribution. Contrast is computed as the root mean square contrast of the samples, meaning the standard deviation of intensities after greyscale conversion, similar to [34, 38].

SWG implementation details and pseudocode. To make SWG compatible with DiT, we add positional encoding interpolation on DiT-XL to process various resolutions. The choice of $k \approx \frac{5}{8}H$ is determined by practical considerations of Unets. Specifically, k must satisfy $\frac{k}{2^n} \in \mathbb{N}$, where n is the number of downsampling steps of the Unet, and the diffusion model $\boldsymbol{\epsilon}_{\text{pos}}$ must be able to process multiple image resolutions. We provide an implementation of SWG in the Section App. I.

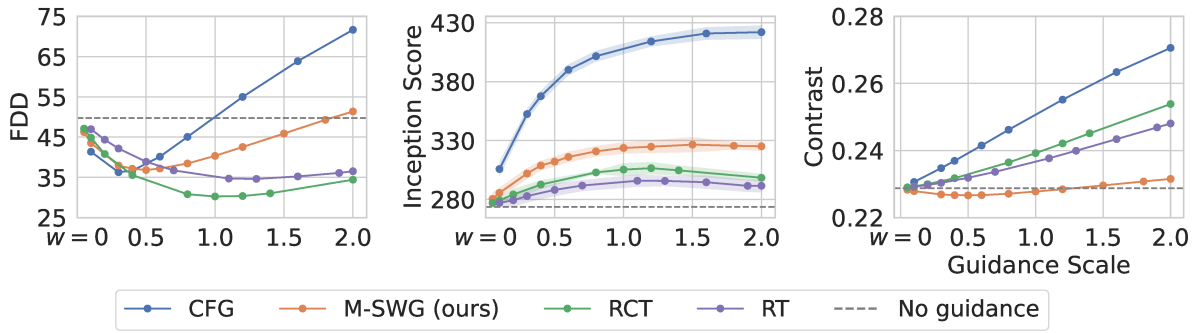


Figure 9.7: **Visualizing the achieved tradeoff by varying the guidance scale w (x-axis) for FDD (↓, left), Inception score (IS ↑, center), and contrast (↓, right) using EDM2-XXL**. M-SWG does **not** exhibit undesired overcontrast compared to other state-of-the-art methods.

```python
import torch

def SWG(x, e_pos, t, labels, N, k, w):
  # x: noisy images of size [bs, c, H, H]
  # e_pos: network i.e. Unet
  # t: timestep
  # labels: class conditions
  # N: number of crops in [4,9,16,...]
  # k: crop size
  # w: guidance scale

  steps_per_dim = int(torch.sqrt(N))
  y_pos = e_pos(x, t, labels)
  bs, c, H, _ = x.shape
  y_neg = torch.zeros_like(x)
  overlap = torch.zeros_like(x)
  stride = (H - k) // (steps_per_dim - 1)
  for i in range(steps_per_dim):
    for j in range(steps_per_dim):
      # left edge (le), right edge (re)
      le, re = stride * i, stride * i + k
      # top edge (te), bottom edge (be)
      te, be = stride * j, stride * j + k
      # x_p is of shape [bc, c, k, k]
      x_p = x[:, :, te:be, le:re]
      y_crop = e_pos(x_p, t, labels)
      y_neg[:, :, te:be, le:re] += y_crop
      o_p = torch.ones_like(y_crop)
      overlap[:, :, te:be, le:re] += o_p
  # pixel averaging
  y_neg = y_neg / overlap
  # Equation 2 in the main text
  return y_pos + w * (y_pos - y_neg)

```

Figure 9.8: Example pseudocode for sliding window guidance (SWG) in Python using the torch library.

10. Conclusion and Publication List

Visual representation learning using foundational models in computer vision has led to crucial advancements. However, adapting these models to settings with little or no labeled data remains challenging. This thesis contributes new methods that repurpose both visual and multimodal foundation models on three tasks: a) image clustering, b) visual out-of-distribution detection, and c) label-free generative modeling. Below we summarise our contributions per chapter in Part II:

In *Chapter 4*, we focused on how to perform image clustering using a pretrained foundational model. It describes a novel clustering method, called TEMI, with strong outperformance over naively applying k-means on image features. Similar to contrastive learning, TEMI is inspired by information theory and specifically pointwise mutual information. TEMI involves training a small clustering head using nearest neighbors in feature space as positive pairs. The head is trained to assign positive pairs in the same cluster using a novel objective that has a single bounded hyperparameter, which requires minimal to no tuning. This pipeline significantly simplifies previous deep learning-based methods while it outperforms classical methods such as k-means on several vision benchmarks, including ImageNet-1K.

In *Chapter 5*, we focused on how to apply contrastive language-image pretrained models for visual OOD detection. A holistic study is presented for several scenarios based on the availability of ground truth labels. There, we developed a new method called pseudo-label-probing (PLP), which leverages only the ground truth label names of the in-distribution. Instead of simply computing top-1 image-text similarities as the anomaly score, PLP first predicts pseudo-labels per sample. Then, a linear layer is trained on these pseudo-labels, and the energy OOD detection score is computed. In contrast to previous methods, no tuning of the visual encoder is required, while PLP scales better than the compared baselines with respect to the model size up to 2B parameters.

In *Chapter 6*, we focus on finding alternative conditioning for unconditional image synthesis. We propose using cluster assignments from learned features as conditioning in generative diffusion models for unlabelled image data. Moreover, a strategy for estimating the optimal range of clusters for generative learning is presented. Experimental results using the EDM diffusion model baseline on CIFAR10, CIFAR100, and FFHQ demonstrate a threefold training sampling efficiency compared to unconditional synthesis, with performance (FID) comparable to ground-truth class-conditioned models.

In *Chapter 7*, we focus on developing challenging image clustering benchmarks based on ImageNet-21K. To this end, we conduct a comprehensive study of feature-based clustering methods at scale, disentangling the impact of the following data-related factors: i) class imbalance, ii) class granularity, iii) the presence of easily recognizable classes, and iv) the ability to capture multiple classes. We developed a novel approach for reannotating GT labels to more fine-grained classes using CLIP in conjunction with the WordNet

semantic tree.

In *Chapter 8*, we focused on OOD detection *without* access to ground truth in-distribution label names. Instead, the label names that likely correspond to the in-distribution (positives) are mined from a large and diverse text corpus. We propose a novel OOD detection method that utilizes text corpora for positive label mining, called ClusterMine. ClusterMine is the first method to achieve state-of-the-art OOD detection performance on ImageNet-1K benchmarks without access to the ground truth in-distribution label names. ClusterMine combines visual-only sample consistency (via clustering) and zero-shot image-text consistency to decide which label names correspond to the in-distribution. Experimental results reveal that ClusterMine achieves state-of-the-art performance in full-spectrum OOD detection.

In *Chapter 9*, we focus on understanding and developing sampling methods for diffusion models. We proposed a new method, called masked sliding window guidance (M-SWG), that upweights long-range spatial dependencies. Precisely, M-SWG guides the diffusion model with itself by selectively restricting its receptive field using crops. M-SWG can be applied as an alternative to other sampling methods (standalone) or it can be combined with existing sampling methods. As a standalone sampling method, M-SWG achieves a superior Inception score (IS) compared to previous state-of-the-art training-free approaches, without introducing sample oversaturation. In conjunction with existing guidance methods, such as classifier-free guidance (CFG), M-SWG reaches state-of-the-art Frechet distance on various ImageNet-1K benchmarks. Finally, M-SWG is applicable to unconditional image synthesis with significant improvements over unconditional generation.

Publication List

Below we present the author's full publication list during his PhD candidacy in chronological order based on the first submission date of each paper:

1. Rafiee, N., Gholamipoor, R., **Adaloglou, N.**, Jaxy, S., Ramakers, J., & Kollmann, M. (2022, September). Self-supervised anomaly detection by self-distillation and negative sampling. In International Conference on Artificial Neural Networks (pp. 459-470). Cham: Springer Nature Switzerland.

2. **Adaloglou*, N.**, Michels*, F., Kalisch, H., & Kollmann, M. (2023). Exploring the limits of deep image clustering using pretrained models. In 34th British Machine Vision Conference 2023 (BMVC 2023). BMVA.

3. Kim*, V., **Adaloglou*, N.**, Osterland, M., Morelli, F. M., Halawa, M., König, T., ... & Marin Zapata, P. A. (2025). Self-supervision advances morphological profiling by unlocking powerful image representations. Scientific Reports, 15(1), 4876.

4. **Adaloglou, N.**, Michels, F., Kaiser, T., & Kollmann, M. (2024). Adapting Contrastive Language-Image Pretrained (CLIP) Models for Out-of-Distribution Detection. Trans. Mach. Learn. Res.

5. **Adaloglou, N.**, Kaiser, T., Michels, F., & Kollmann, M. (2025). Rethinking cluster-conditioned diffusion models for label-free image synthesis. In 2025 IEEE/CVF

Winter Conference on Applications of Computer Vision (WACV) (pp. 3603-3613). IEEE.

6. **Adaloglou, N.**, Michels, F., Senft*, K., Petrusheva*, D., & Kollmann, M. (2024). Scaling up deep clustering methods beyond imagenet-1k. arXiv preprint arXiv:2406.01203.

7. **Adaloglou, N.**, Kaiser, T., Iagudin, D., & Kollmann, M. (2025). Guiding a diffusion model with itself using sliding windows. In 36th British Machine Vision Conference (BMVC 2025). BMVA.

8. **Adaloglou, N.**, Petrusheva*, D., Asker*, M., Michels, F., & Kollmann, M. (2026). ClusterMine: Robust label-free visual out-of-distribution detection via concept mining from text corpora. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV).