Journal of Machine Learning Research 27 (2026) 1-46 Submitted 3/25; Revised 8/26; Published 8/26

A Theoretical Framework for Masked Pretraining (MPT)

Qi Zhang [1] * ZHANGQ327@STU.PKU.EDU.CN
Runyu Zhou [1] * RUNYUZHOU25@STU.PKU.EDU.CN
Yifei Wang [2] * YIFEI_W@MIT.EDU
Yisen Wang [1] † YISEN.WANG@PKU.EDU.CN

[1] *State Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University*
[2] *CSAIL, MIT*

Editor: Sanjiv Kumar

Abstract

Recently, Masked Pretraining (MPT) based on reconstruction pretraining tasks has risen to a promising self-supervised learning paradigm across various domains and achieves remarkable performance in multiple downstream tasks. However, the theoretical understanding of the working mechanism behind MPT is still limited. In this paper, we introduce a new theoretical framework to analyze MPT and understand the crucial role of masking in extracting meaningful representations. We establish theoretical connections between MPT and another popular self-supervised paradigm: contrastive learning. We prove that the masking technique implicitly creates positive pairs that are semantically similar and the reconstruction loss pulls them together in the feature space. Besides, as a result of the implicit alignment, we point out the dimensional collapse issue of MPT and propose a Uniformity-enhanced MPT (U-MPT) loss that can effectively address this issue and bring significant improvements in downstream tasks including linear evaluation, cross-dataset fine-tuning and out-of-distribution generalization on real-world data sets. Furthermore, we establish downstream guarantees of U-MPT and theoretically analyze the influence of masking strategies. Based on the theoretical analysis, we propose a new masking strategy which enhances the downstream performance of MPT and explains current improvements of masking strategies with our theoretical perspective.

Keywords: masked pretraining, dimensional collapse, masking strategy, representation learning theory

1. Introduction

In recent years, self-supervised learning (SSL) has emerged as a promising paradigm for learning meaningful representations without the need for labeled data. SSL methods can be broadly categorized into two types: contrastive pretraining and masked pretraining. Contrastive pretraining methods (Chen et al., 2020a; He et al., 2020) apply data augmentation on natural samples and train neural networks by aligning different augmented views

∗. Equal Contribution
†. Corresponding Author

of the same sample in the feature space. On the other hand, masked pretraining methods (Devlin et al., 2019; He et al., 2022) apply masks to samples and predict the masked portions based on the unmasked views. Although contrastive and masked pretraining differ in their objectives, both approaches have demonstrated impressive performance across various domains, including vision (He et al., 2022; Chen et al., 2020a), language (Devlin et al., 2019; Zhang et al., 2022b), and multi-modal tasks (Radford et al., 2021; Khan and Fu, 2021; Zhang et al., 2023).

Due to the remarkable empirical success of self-supervised learning, researchers try to theoretically analyze the working mechanism behind these paradigms. Various theoretical frameworks have been proposed to analyze contrastive pretraining from different perspectives. For example, Saunshi et al. (2019) establish the connection between the pretraining contrastive objective and downstream classification error. HaoChen et al. (2021) observe that the contrastive loss is equal to a matrix decomposition objective, and Hjelm et al. (2019) understand the contrastive pretraining process with the information theory. However, the theoretical analysis of masked pretraining is still relatively underexplored.

In this paper, we introduce a new theoretical framework for masked pretraining (MPT). Taking the representative masked pretraining method Masked Autoencoder (MAE) (He et al., 2022) as an example, we note that there are two core components: autoencoders and masking. As vanilla autoencoders exhibit much inferior performance than contrastive paradigms (Chen et al., 2020a), there exists a common belief that masking is the key factor for the success of MPT. To understand the role of masking in MPT, we first establish a close connection between contrastive and masked pretraining. Specifically, we prove that the masking technique creates implicit positive pairs that are semantically similar and the reconstruction loss pulls these positive pairs together. In contrastive pretraining, merely aligning positive pairs can lead to feature collapse, as the alignment loss can be minimized by mapping different inputs to a constant value. Given the connection between contrastive and masked pretraining, we wonder whether MPT also faces this collapse issue. Our findings show that while MPT avoids the trivial constant solution, it still suffers from dimensional collapse, where the effective dimensions of the learned representations are quite limited. To address this problem, we draw inspiration from contrastive learning and add a uniformity regularizer into the MPT objective. Specifically, we propose Uniformity-enhanced MPT (U-MPT), which calculates the feature similarity between independent samples and encourages them to be pushed apart. Empirically, we observe that U-MPT effectively resolves the dimensional collapse and significantly boosts MPT performance on downstream linear evaluation, cross-dataset fine-tuning and out-of-distribution generalization tasks.

Built on the connection between contrastive pretraining and masked pretraining, we then establish theoretical guarantees for U-MPT's downstream generalization performance. Our theoretical bounds demonstrate that the generalization performance of MPT is decided by two factors: first, the unmasked views should preserve the core semantics of natural samples (like label information), and second, semantically similar unmasked views should be selected as an implicit positive pair. With a toy model, we prove that a lower mask ratio helps the unmasked views retain label information while a higher mask ratio increases the probability that semantically similar unmasked views are selected as implicit positive pairs. This introduces a trade-off in selecting the optimal mask ratio. Based on this insight, we propose a surrogate metric to estimate downstream guarantees on real-world data sets,

and we confirm that it can reliably estimate the optimal mask ratio in practice (75%). Furthermore, with the theoretical analysis of masking operation in MPT, we propose an improved masking strategy, which exhibits better downstream classification performance than random masking. Meanwhile, we observe that many of the recent advancements in masking strategies and objective designs can be analyzed and explained through our theoretical framework, which further verifies the generality of our proposed analysis. The contents of the paper are organized as follows:

- In Section 2, we summarize the related work of understanding self-supervised learning and further introduce the mathematical problem setup in Section 3.
- In Section 4, we establish theoretical connections between masked pretraining and contrastive pretraining, showing that the masking operation implicitly creates positive pairs that are semantically similar and the reconstruction objectives align them in the feature space.
- In Section 5, we propose the uniformity-enhanced MPT (U-MPT) objective, which alleviates the dimensional collapse issue in MPT and significantly improves the downstream performance of MPT in downstream tasks including linear evaluation, cross-dataset fine-tuning, and out-of-distribution generalization on real-world data sets.
- In Section 6, we establish theoretical guarantees on the downstream performance of MPT. With a toy model, we theoretically analyze the influence of mask ratio on the downstream error of MPT. Furthermore, we propose a metric to estimate MPT's downstream performance, which accurately predicts the optimal mask ratio.
- In Section 7, we improve the masking strategy in MPT based on our theoretical analysis. Besides, we also find that many of the current improvements can be explained with our theoretical framework.

Remark. A preliminary work was published at NeurIPS 2022 as a Spotlight paper (Zhang et al., 2022a), while we add substantially new results, both theoretically and empirically, in this manuscript, aiming to provide a more comprehensive and in-depth analysis of the generalization property of masked pretraining. Specifically, we add the following key results:

- In Sections 3 and 4, we incorporate cross-entropy (CE) loss-based MPT methods into our theoretical framework alongside the reconstruction loss-based approaches, thereby enhancing the comprehensiveness of our analysis.
- In Section 5, we verify our proposed U-MPT objectives in more downstream scenarios, including cross-dataset fine-tuning and out-of-distribution generalization, which further verify the effectiveness and advantages of U-MPT.
- In Section 6, we theoretically analyze the influence of mask ratio on downstream guarantees with a toy model. We also propose a new metric to estimate the theoretical bounds.
- In Section 7, we propose new masking and reconstruction strategies based on our theoretical analysis. Besides, we explain the current improvements of MPT with our theoretical framework.

2. Related Work

In this section, we review two lines of work most relevant to our study: the empirical development and the theoretical efforts of masked pretraining (MPT) methods.

Masked Pretraining Methods in Practice. Deep supervised learning algorithms typically require extensive labeled data to achieve strong performance, yet manual annotation is costly and labor-intensive. Self-supervised learning (SSL) addresses this challenge by learning discriminative representations directly from unlabeled data. Among SSL paradigms, contrastive learning (CL) and Masked Pretraining (MPT) approaches have emerged as dominant frameworks. CL trains models to pull semantically similar samples together in the feature space while pushing dissimilar samples away, achieving notable success in vision and language tasks (Chen et al., 2020a; He et al., 2020; Grill et al., 2020; Wang et al., 2021; Chen and He, 2021; Zbontar et al., 2021; Bardes et al., 2022; Cui et al., 2023; Wang et al., 2023b). MPT methods based on reconstruction tasks have also demonstrated impressive performance (Devlin et al., 2019; He et al., 2022; Du et al., 2024). During the pretraining process, MPT methods reconstruct masked regions of input data to learn transferable representations. Key MPT methods include BERT (Devlin et al., 2019), BeiT (Bao et al., 2022), SimMIM (Xie et al., 2022), and MAE (He et al., 2022), which differ in reconstruction targets: BEiT and BERT optimize token-level cross-entropy losses, while SimMIM and MAE minimize pixel-wise losses.

Theoretical Understandings of Masked Pretraining Methods. The empirical success of SSL has made it important to theoretically understand its underlying mechanisms. For contrastive learning, Saunshi et al. (2019) establish the downstream guarantee of contrastive learning representations. Wang et al. (2022); Zhang et al. (2025b) revise their bounds by resolving the class collision issues, and develop a new understanding of contrastive learning from the perspective of augmentation overlap. Wang and Isola (2020) identify alignment and uniformity as two key properties of contrastive learning. HaoChen et al. (2021) introduce the augmentation graph framework to characterize the downstream performance of spectral contrastive solutions. Cui et al. (2025) propose an augmentation-aware error bound which enables the explanation of specific types of data augmentation in contrastive learning. Zhang et al. (2023) establish generalization guarantees for multimodal contrastive learning. However, compared to contrastive learning, the theoretical analysis of MPT methods remains relatively limited. Cao et al. (2022) analyze the attention mechanism of MPT through an integral kernel perspective. Kong et al. (2023) formulate the underlying data-generating process in MPT as a hierarchical latent variable model. Liu et al. (2022) study masked prediction from a parameter identifiability perspective and characterize when the underlying parameters can be recovered from masked prediction objectives. Meng et al. (2024) reveal representation deficiency in masked language modeling and theoretically show that the mask token can occupy representation capacity, leading to rank deficiency in learned representations. Li et al. (2024) establish a theoretical framework for generative masked language modeling and analyze its statistical properties, including the impact of masking ratios on sample complexity and learning efficiency. Zhang et al. (2024) theoretically compare masked and autoregressive pretraining, showing that the flexible prediction targets in masked pretraining induce richer inter-sample connections and lead to advantages in downstream classification. However, most of these studies predominantly analyze model

architecture and ignore the core designs of MPT methods, i.e., the masking technique. Our work bridges this gap by formalizing the role of masking and deriving the first theoretical guarantees for MPT's downstream performance. We note that while both HaoChen et al. (2021) and our work adopt a graph-based perspective, the corresponding graphs characterize different learning mechanisms. Their augmentation graph defines edges through data augmentations and is developed for analyzing contrastive learning, whereas our mask graph defines edges through reconstruction relationships between masked and unmasked views and is used to characterize the implicit alignment induced by masked reconstruction.

3. Mathematical Formulation for Different Masked Pretraining Methods

In this section, we introduce mathematical formulations of Masked Pretraining (MPT) approaches. We categorize different MPT approaches based on the type of loss used, i.e., the pixel-wise reconstruction loss and the cross-entropy loss. We start by introducing some common definitions for clarity.

Let $\bar{x}$ denote a natural sample from a data set $\mathcal{D}$. In computer vision (CV), $\bar{x}$ typically refers to an image, while in natural language processing (NLP), it corresponds to a sequence of words. The sequence $\bar{x}$ is divided into K elements, where each element is either a patch of size s (e.g., 16×16 in ViT) in CV, such that $\bar{x} \in \mathbb{R}^{K \times s}$, or a token in NLP. To generate a masked view, a random binary mask $m \in \{0,1\}^K$ is applied, where each entry m_k is set to 0 with probability ρ (the mask ratio). This process masks a fraction ρ of the elements, resulting in two complementary views of the sequence: the unmasked view x_1 and the masked view x_2. These views are formally defined as:

$$x_1 = \bar{x}[m], \quad x_2 = \bar{x}[1-m], \tag{1}$$

where $\bar{x}[m]$ selects the elements of $\bar{x}$ corresponding to the positions where $m_k = 1$. Let n_1 and n_2 denote the number of unmasked and masked elements. By construction, $n_1 + n_2 = K$, with $n_1 = (1-\rho)K$ and $n_2 = \rho K$.

The Mask Graph $\mathcal{G}_M$ of MPT. We notice that MPT essentially learns to pair the unmasked and masked views x_1, x_2 via the reconstruction task, which can be characterized by a *mask graph* $\mathcal{G}_M$. Denote the set of all unmasked views as $\mathcal{X}_1 = \{x_1\}$ and the set of all masked views as $\mathcal{X}_2 = \{x_2\}$, where the two sets are assumed to be finite (can be exponentially large), *i.e.,* $|\mathcal{X}_1| = N_1, |\mathcal{X}_2| = N_2$. The mask graph $\mathcal{G}_M$ over the joint set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ is defined here:

- Node: each view $x \in \mathcal{X}$.
- Edge: the edge weight w_{x_1,x_2} between any $x_1, x_2 \in \mathcal{X}$ is defined as their joint probability $\mathcal{M}(x_1, x_2) = \mathbb{E}_{\bar{x}} \mathcal{M}(x_1, x_2 | \bar{x})$. In other words, there is an edge between two views (*i.e.,* $w_{x_1,x_2} > 0$) if and only if they are complementary views generated by masking.

Considering the masking process in Eq. 1, when the masking ratio $\rho \neq 0.5$, the two sets $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint by construction. Therefore, there only exist edges *between* the two sets and there is no edge *within* either set. Thus, the mask graph $\mathcal{G}_M$ is a *bipartite* graph in this case. Its adjacency matrix can be defined as $A_M \in \mathbb{R}^{N_2 \times N_1}$ where $(A_M)_{x_2,x_1} = w_{x_1,x_2}$ for $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$. As long as $\rho \neq 0.5$, $N_1 \neq N_2$ and A_M is not necessarily a square matrix.

We can define the normalized adjacency matrix as $\bar{A}_M = D_2^{-1/2} A D_1^{-1/2}$, where D_1, D_2 are the diagonal degree matrices with elements $d_{x_1} = \sum_{x_2} w_{x_1,x_2}$ and $d_{x_2} = \sum_{x_1} w_{x_1,x_2}$, respectively.

3.1 Methods with Pixel-wise Reconstruction Loss

A wide range of MPT methods in CV is based on pixel-wise reconstruction loss, which measures differences between the original images and the reconstructed regions (He et al., 2022; Xie et al., 2022). For instance, MAE (He et al., 2022) employs an encoder-decoder architecture $h = g \circ f$, where f is the encoder that maps unmasked patches x_1 to latent features z_1, and g is the decoder that reconstructs the complementary masked view x_2 by mapping z_1 and the mask token back to the pixel space. During the pretraining process, MAE minimizes the l_2-norm loss between the original and reconstructed masked areas:

$$\mathcal{L}_{\text{MAE}}(h) = \mathbb{E}_{\bar{x}} \mathbb{E}_{x_1, x_2 | \bar{x}} \left\| g(f(x_1)) - \hat{x}_2 \right\|_2^2, \tag{2}$$

where $\hat{x}_2$ is the l_2-normalized version of x_2.

SimMIM (Xie et al., 2022) also uses an encoder-decoder framework, where the encoder f takes the unmasked view x_1 and a mask token (spread over the masked positions) as input, outputting the corresponding features. A linear layer g is then used to reconstruct the masked patches from these features, and the loss function is defined as an l_1 reconstruction loss:

$$\mathcal{L}_{\text{SimMIM}}(h) = \frac{1}{n_2} \mathbb{E}_{\bar{x}} \mathbb{E}_{x_1, x_2 | \bar{x}} \left\| g(f(x_1)) - x_2 \right\|_1. \tag{3}$$

3.2 Methods with Cross-Entropy Loss

While many MPT methods in the vision domain rely on the pixel-wise reconstruction loss (such as l_1 or l_2 loss), MPT pretraining approaches usually adopt the cross-entropy (CE) loss in the language domains (Devlin et al., 2019). Taking the popular framework BERT (Devlin et al., 2019) as an example, the pretraining objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{\text{BERT}} = -\mathbb{E}_{x_1, x_2} \sum_k \log \Pr(x_{2,k}^+ \mid x_1), \tag{4}$$

where x_1 and x_2 are the unmasked and masked portions of a sequence from a language data set, $x_{2,k}^+$ denotes the k-th token of the masked view x_2. This formulation captures the likelihood of predicting the masked tokens in x_2 based on the unmasked portion x_1. Specifically, the predicted probability of a token $x_{2,k}^+$ in the masked portion can be formulated as:

$$\mathcal{L}_{\text{BERT}} = -\mathbb{E}_{x_1, x_2} \sum_k \log \frac{\exp(f(x_1)^\top w_{x_{2,k}^+})}{\sum_{x_{2,k}^-} \exp(f(x_1)^\top w_{x_{2,k}^-})}, \tag{5}$$

where $x_{2,k}^-$ denotes irrelevant tokens in the vocabulary, $f(x_1)$ is the encoder output given the input x_1, and w_i is the feature embedding of the i-th token in the weight matrix of the token classification head.

Inspired by the impressive performance of BERT in the language domain, Bao et al. (2022) adapts the concept of masked token prediction to vision tasks. Unlike methods that operate directly on raw pixels, BEiT employs a discrete variational autoencoder (dVAE) to transform an image $\bar{x}$ into a sequence of discrete tokens. Specifically, each patch of the image is tokenized into an index from a visual codebook. BEiT then learns to predict the token indices of the masked patches using the embeddings of the visible regions by minimizing the following loss:

$$\mathcal{L}_{\text{BEiT}} = -\mathbb{E}_{x_1,x_2} \sum_k \log \Pr(x_{2,k}^+ \mid x_1),$$

where x_1 and x_2 represent the unmasked and masked views of the image, $x_{2,k}^+$ denotes the k-th token of the masked view x_2.

These reconstruction-based methods, though differing in the specific loss functions used, share the common goal of predicting masked content from incomplete inputs. In the next section, we will reveal their intrinsic connections in the lens of sample alignment.

4. Masked Pretraining Performs Implicit Contrastive Learning

Masked pretraining (MPT) and contrastive learning (CL) are two of the most widely used self-supervised paradigms and both of them exhibit remarkable performance across various downstream tasks (Chen et al., 2020a; He et al., 2022). While contrastive learning has been extensively studied with several theoretical frameworks analyzing its mechanisms from different perspectives (Saunshi et al., 2019; HaoChen et al., 2021; Hjelm et al., 2019), the theoretical exploration of MPT approaches remains relatively underdeveloped. In this paper, we provide a theoretical analysis of MPT by establishing a fundamental connection between MPT and CL. Specifically, we prove that MPT implicitly aligns semantically similar pairs through its masking mechanism, similar to the alignment between positive pairs in contrastive methods. In Section 4.1, we formalize this relationship, proving that minimizing the pixel-wise MPT reconstruction loss enforces alignment between semantically similar samples. Furthermore, in Section 4.2, we extend this analysis to cross-entropy-based MPT methods (e.g., BEiT, BERT), showing that their loss functions implicitly align encoder outputs with token embeddings, mirroring contrastive objectives.

4.1 Pixel-wise MPT Implicitly Aligns Positive Input Pairs

Masked pretraining approaches typically involve an encoder-decoder architecture, similar to that of autoencoders. In this context, we start by assuming that the encoder-decoder structure in MPT can perform the fundamental task of a standard autoencoder, *i.e.,* reconstructing the original input.

Assumption 1. *For any decoder g, let $\mathcal{F}_g \subseteq \mathcal{F}$ denote a non-empty family of pseudo-inverse encoders associated with g. We define the best achievable autoencoding error for g as*

$$\varepsilon(g) = \min_{f \in \mathcal{F}_g} \mathbb{E}_x \|g \circ f(x) - x\|_2^2,$$

where x represents either unmasked data x_1 or masked data x_2. We assume that the minimum is attained by some $f_g \in \mathcal{F}_g$, i.e.,

$$f_g \in \arg\min_{f \in \mathcal{F}_g} \mathbb{E}_x \|g \circ f(x) - x\|_2^2.$$

The assumption characterizes the reconstructability of a given decoder g through its associated pseudo-inverse encoders, with $\varepsilon(g)$ quantifying the best achievable autoencoding error. For sufficiently expressive decoders, such reconstructability is plausible in practice, given the strong empirical performance of deep networks on autoencoding tasks (Kingma and Welling, 2014; Jing et al., 2020). Moreover, the Transformer architectures commonly used in MPT possess universal approximation capabilities (Yun et al., 2020), further supporting the existence of encoder-decoder pairs with small autoencoding error. We emphasize, however, that $\varepsilon(g)$ is decoder-dependent and need not be uniformly small for all g.

However, it is worth noting that the vanilla autoencoder task fails to learn representations as meaningful as those achieved by MPT. For instance, when no masks are applied, the linear probing accuracy of MAE has a significant decline from 61.2% to 17.4% on ImageNet-100. This outcome suggests that the autoencoding ability alone, as explored in previous studies (Cao et al., 2022), is insufficient to explain the effectiveness of MPT. Hence, this motivates us to delve further into the masking mechanism employed by MPT.

MPT Performs Asymmetric Input-Output Alignment on the Mask Graph. For the sake of simplicity, we assume that MPT employs an l_2 reconstruction loss. First, we show that the MPT loss can be lower bounded by an *asymmetric* alignment loss between the two complementary views x_1, x_2 using two-branch autoencoders $h = g \circ f$ (real) and $h_g = g \circ f_g$ (pseudo), respectively.

Theorem 2. *Under Assumption 1, the MPT loss can be lower bounded by*

$$\mathcal{L}_{\text{MPT}}(h) \geq \mathcal{L}_{\text{asym}}(h) - \varepsilon(g) + \text{const}, \tag{6}$$

$$\text{and } \mathcal{L}_{\text{asym}}(h) = -\mathbb{E}_{x_1,x_2} h(x_1)^\top h_g(x_2) = -\operatorname{tr}(H_g^\top \bar{A}_M H), \tag{7}$$

where H denotes the output matrix of h on $\mathcal{X}_1$ whose x_1-th row is $H_{x_1} = \sqrt{d_{x_1}} h(x_1)$, and H_g denotes the output matrix of h_g on $\mathcal{X}_2$ whose x_2-th row is $(H_g)_{x_2} = \sqrt{d_{x_2}} h_g(x_2)$.

Consequently, when the MPT loss is small, it implies a reduced alignment loss (as a lower bound). This indicates that MPT, with its encoder-decoder architecture, implicitly aligns the masked and unmasked views. However, it is still unclear why aligning these two complementary views contributes to learning meaningful features. To gain a deeper understanding, we further explore the effect of masking in MPT. We find that via masking, MPT generates *implicit connections* among different *input samples* in the form of *2-hop*

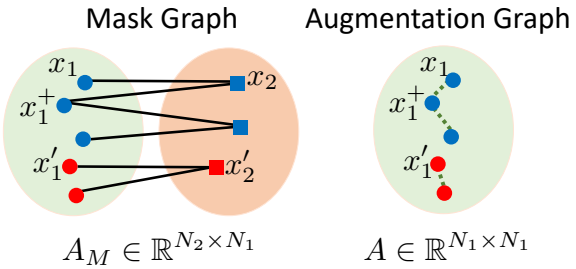


Figure 1: An illustration of the mask graph and the corresponding augmentation graph of MPT. Different colors denote different belonging classes.

connectivity. Consider a pair of 2-hop input neighbors, denoted as x_1 and x_1^+, both belonging to $\mathcal{X}_1$, and sharing a common complementary target view, $x_2 \in \mathcal{X}_2$ (more likely to occur under a larger mask ratio ρ), as illustrated in Figure 1. By enforcing x_1 and x_1^+ to reconstruct the same output x_2, MPT implicitly maps their features together. In this way, the 2-hop input neighbors serve as *positive pairs* that are implicitly aligned as in contrastive learning.

Motivated by this observation, we seek to establish a formal connection between MPT and contrastive learning below. Specifically, we will show that similar to the data augmentations employed in contrastive learning, the masking mechanism in MPT also introduces implicit affinities between *input samples* and creates (implicit) *positive pairs*. To present this connection formally, we introduce an augmentation graph $\mathcal{G}_A$, which models the relationship between all input samples in $\mathcal{X}_1$. It is important to note that this augmentation graph $\mathcal{G}_A$ is distinct from the mask graph $\mathcal{G}_M$, which captures the input-output relationship between $\mathcal{X}_1$ and $\mathcal{X}_2$.

The Augmentation Graph $\mathcal{G}_A$ of MPT. To model the mask-induced affinity between all unmasked views in $\mathcal{X}_1$, we construct an augmentation graph $\mathcal{G}_A$.[1] Specifically, for any pair of views x_1 and x_1' in $\mathcal{X}_1$, we define the edge weight in the adjacency matrix A as the probability of having the same target view, *i.e.,* $\mathcal{A}(x_1, x_1') = \mathbb{E}_{x_2}[\mathcal{M}(x_1|x_2)\mathcal{M}(x_1'|x_2)]$.[2] With this formulation, we define a *symmetric* alignment loss with respect to the positive pairs $(x_1, x_1^+) \sim \mathcal{A}(x_1, x_1^+)$ drawn according to the augmentation graph:

$$\mathcal{L}_{\text{align}}(h) = -\mathbb{E}_{x_1, x_1^+} h(x_1)^\top h(x_1^+). \tag{8}$$

MPT Performs Symmetric Input Alignment on the Augmentation Graph. Built upon the augmentation graph constructed above, we theoretically verify the intuition on 2-hop connectivity. To achieve this, we establish a relationship between the asymmetric input-output alignment loss on the mask graph and the symmetric input alignment loss on the augmentation graph in the following theorem.

Theorem 3. *The asymmetric alignment loss on the mask graph (Eq. 7) can be lower bounded by the symmetric alignment loss on the augmentation graph (Eq. 8):*

$$\mathcal{L}_{\text{asym}}(h) \geq \frac{1}{2}\mathcal{L}_{\text{align}}(h) + \text{const}. \tag{9}$$

Proof Sketch. We provide a proof sketch of this inequality as it is the key to our analysis. We first symmetrize the asymmetric alignment loss $\mathcal{L}_{\text{asym}}(h)$ with an arithmetic inequality, and then establish its equivalence to the symmetric alignment loss $\mathcal{L}_{\text{align}}(h)$. The derivation highlights the intrinsic connection between the mask graph (adjacency matrix A_M) and the

1. We can also construct an augmentation graph $\mathcal{G}_A'$ on the output space $\mathcal{X}_2$, where the (x_2, x_2')-th element of the adjacency matrix A' is defined by $\mathcal{A}(x_2, x_2') = \mathbb{E}_{x_1}[\mathcal{M}(x_2|x_1)\mathcal{M}(x_2'|x_1)]$. As the results are quite similar, we mainly take the input space as an example in this work.
2. $\mathcal{M}(x_1|x_2) = \mathcal{M}(x_1, x_2)/\mathcal{M}(x_2)$ can further be calculated by marginalizing $\mathcal{M}(x_1, x_2|\bar{x})$.

augmentation graph (adjacency matrix A).

$$
\begin{aligned}
\mathcal{L}_{\text{asym}}(h) &= \mathbb{E}_{x_1,x_2} h(x_1)^\top h_g(x_2) \\
&= -\operatorname{tr}(H_g^\top \bar{A}_M H) && \text{(reformulated to the mask graph (Theorem 2))} \\
&\geq -\frac{1}{2}(\|\bar{A}_M H\|^2 + \|H_g\|^2) && \text{(because } \operatorname{tr}(AB) \leq \frac{1}{2}(\|A\|^2 + \|B\|^2)\text{)} \\
&= -\frac{1}{2}\operatorname{tr}(H^\top \bar{A}_M^\top \bar{A}_M H) - \frac{1}{2} && \text{(because } \|H_g\|^2 = \sum_{x_2} d_{x_2} \|h_g(x_2)\|^2 = 1\text{)} \\
&= -\frac{1}{2}\sum_{x_1,x_1'} A_{x_1,x_1'} h(x_1)^\top h(x_1') - \frac{1}{2} && \text{(transformed to the augmentation graph)} \\
&= \frac{1}{2}\mathcal{L}_{\text{align}}(h) - \frac{1}{2}. && \text{(following the definition in Eq. 8)}
\end{aligned}
$$

Combining Theorem 2 and Theorem 3, we derive the main theorem, which shows that MPT's reconstruction loss can be bounded by the symmetric alignment loss of the positive input pairs (x_1, x_1^+) drawn according to the augmentation graph. We provide visualization examples of the augmentation graph in Appendix C.

Theorem 4. *Under Assumption 1, MPT's reconstruction loss (Eq. 2) can be lower bounded by the alignment loss between positive pairs $(x_1, x_1^+) \sim \mathcal{A}(x_1, x_1^+)$,*

$$
\mathcal{L}_{\text{MPT}}(h) \geq \frac{1}{2}\mathcal{L}_{\text{align}}(h) - \varepsilon(g) + \text{const} = -\frac{1}{2}\mathbb{E}_{x_1,x_1^+} h(x_1)^\top h(x_1^+) - \varepsilon(g) + \text{const}. \tag{10}
$$

In this way, we establish a close relationship between the two prominent SSL paradigms (MPT and contrastive learning) by showing that a small MPT loss will imply a small alignment loss of positive input pairs as in contrastive learning. Leveraging this connection, we could establish guarantees for the downstream generalization of MPT (discussed in Section 6).

Comparison to Contrastive Learning. Comparing the alignment loss (Eq. 8) to that of contrastive learning (Chen et al., 2020a; HaoChen et al., 2021), we observe that the main difference lies in that contrastive learning aligns features in the *latent* space of the encoder f, while MPT aligns features in the *output* space with an encoder-decoder architecture $h = g \circ f$. Nevertheless, it is worth noting that most variants of contrastive learning apply a nonlinear projection head g following the encoder f before calculating the alignment loss (Chen et al., 2020a,b; He et al., 2020; Grill et al., 2020; Chen and He, 2021; Ouyang et al., 2025). Similarly, MPT also utilizes a decoder g. Therefore, we may also regard the role of MPT's decoder as the projection head in contrastive learning. Theoretically, if we further assume the bi-Lipschitzness of the decoder, we can show that the MPT loss is further lower bounded by an alignment loss defined in the feature space.

Corollary 5. *Under Assumption 1 and the assumption that the decoder is L-bi-Lipschitz, i.e., $\forall\ (x_1, x_2)$, $\frac{1}{L}\|x_1 - x_2\|^2 \leq \|g(x_1) - g(x_2)\|^2 \leq L\|x_1 - x_2\|^2$. Then, the MPT reconstruction loss can be lower bounded by the alignment loss w.r.t. the encoder outputs:*

$$
\mathcal{L}_{\text{MPT}}(h) \geq -\frac{1}{2L} \cdot \mathbb{E}_{x_1,x_1^+} f(x_1)^\top f(x_1^+) - \varepsilon(g) + \text{const}. \tag{11}
$$

4.2 Cross-Entropy Loss Is Implicitly Equal to Token-Level Alignment

While pixel-wise reconstruction-based MPT methods align features through masked prediction, we further analyze whether cross-entropy-driven approaches like BEiT and BERT also achieve alignment indirectly via token prediction in this section.

Incorporating Eq. (5), the cross-entropy (CE) loss, which is useful in models such as BERT and BEiT, can be uniformly reformulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}} &= -\mathbb{E}_{x_1,x_2}\sum_k \log \Pr(x_{2,k}^+|x_1) \\ &= \sum_k \left[-\mathbb{E}_{x_1,x_2} f(x_1)^\top w_{x_{2,k}^+} + \mathbb{E}_{x_1,x_2}\log\sum_{x_{2,k}^-}\exp(f(x_1)^\top w_{x_{2,k}^-})\right], \end{aligned} \tag{12}$$

where $x_{2,k}^+$ is a ground-truth token and $x_{2,k}^-$ is an irrelevant token. The core objective of CE loss is to maximize the predicted probability of corresponding masked tokens while minimizing the probability of irrelevant tokens. Following the same spirit, we adopt the spectral loss (HaoChen et al., 2021) for simplicity of analysis, i.e.:

$$\mathcal{L}_{\mathrm{spectral}} = -\sum_k \mathbb{E}_{x_1,x_2} f(x_1)^\top w_{x_{2,k}^+} + \sum_k \mathbb{E}_{x_1,x_2,x_{2,k}^-}\left[f(x_1)^\top w_{x_{2,k}^-}\right]^2. \tag{13}$$

From this formulation, we observe that the first term of the cross-entropy loss encourages alignment between the predicted features of the unmasked portion ($f(x_1)$) and the feature vectors of the masked tokens in the classification head ($w_{x_{2,k}^+}$). Consequently, when two unmasked views are aligned with the same masked view, they are also implicitly aligned with each other. As a result, we know that MPT methods with CE loss also implicitly align the unmasked views that share the same masked views.

5. Avoiding Dimensional Collapse in MPT with an Explicit Uniformity Regularization

In Section 4, we have demonstrated that the MPT loss is implicitly equal to an alignment loss. However, in contrastive learning, simply aligning the positive pairs leads to the problem of full feature collapse, because the alignment loss can also be minimized when the encoder produces a constant feature for all inputs. Interestingly, MPT manages to avoid full feature collapse. This raises the question of how MPT achieves this property and we discuss it in this section. Specifically, in Section 5.1, we reveal that the MPT loss can avoid full feature collapse while still suffering from dimensional collapse. In Section 5.2, we propose uniformity-enhanced MPT (U-MPT) loss to solve the dimensional collapse issue. Empirically, we verify the effectiveness of the U-MPT objective in Section 5.3. Furthermore, in Section 5.4, we empirically analyze the advantages of U-MPT and the choices of hyperparameters.

5.1 The Feature Collapse Issue in MPT

As shown in Section 4, during the pretraining process, MPT implicitly aligns the semantically similar images. However, we note that simply aligning the positive pairs may lead to the full feature collapse, i.e., the network can minimize the alignment loss by encoding all input into a constant. To address this issue, various techniques have been proposed in contrastive learning, such as incorporating additional losses to encourage feature uniformity or decorrelation (Chen et al., 2020a; He et al., 2020; Zbontar et al., 2021), or employing asymmetric structural designs (Grill et al., 2020; Chen and He, 2021; Caron et al., 2020). Recent theoretical work further attributes the collapse-avoidance effect of such asymmetric designs to a rank differential mechanism that improves the effective dimensionality of learned representations (Zhuo et al., 2023). However, MPT based on the pixel-wise reconstruction loss avoids this issue without special designs. In the following theorem, we show that minimizing the MPT loss can provably get rid of the full feature collapse.

Theorem 6. *When the encoder fully collapses, i.e., $\forall x \in \mathcal{X}_1, f(x) = c$, the MPT loss has a large lower bound:*

$$\mathcal{L}_{\mathrm{MPT}}(h) \geq \mathrm{Var}(x_2), \tag{14}$$

where $\mathrm{Var}(x_2)$ denotes the variance of masked targets computed on the training data set.

MPT Can Avoid Full Feature Collapse. Since the training data contain diverse images, the variance $\mathrm{Var}(x_2)$ is relatively large. Consequently, unlike the alignment loss in contrastive learning, the MPT loss cannot be minimized (to a small value) by a fully collapsed encoder. The primary reason is that the alignment loss operates in a fully flexible latent space that allows a collapsed encoder to minimize the loss, while in MPT, the reconstruction loss adopts a parameter-invariant and sample-dependent target x_2. As a result, a fully collapsed encoder is unable to minimize the reconstruction loss with respect to x_2. Consequently, MPT is inherently resistant to full feature collapse.

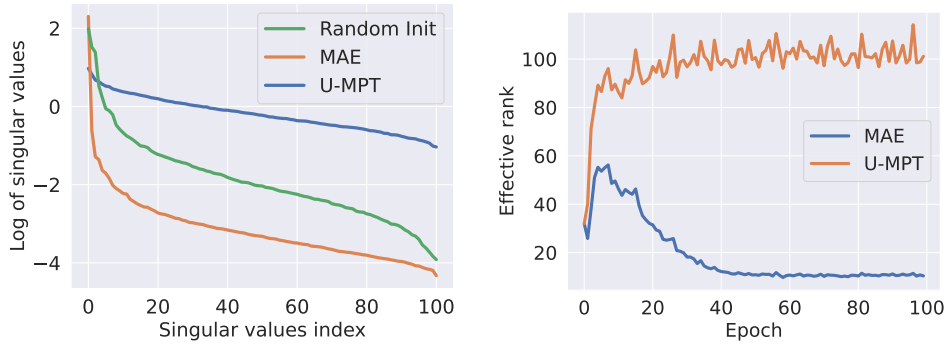
MPT Still Suffers from Dimensional Collapse. Although MPT can prevent full feature collapse, it may still suffer from *dimensional feature collapse* where the learned features lie in a low-dimensional subspace (Hua et al., 2021; Jing et al., 2022), which also limits its representation power. To illustrate that this issue can arise even under optimal reconstruction, we construct a simple example in which perfect masked reconstruction is compatible with a dimensionally collapsed representation.

Proposition 7. *Consider a masked reconstruction problem with a data set of N images, where each image consists of K identical patches:*

$$x_i = [p_i, p_i, \cdots, p_i],$$

where $p_i \in \mathbb{R}^s$ and $p_i \neq p_j$ for any $i \neq j$. Let the representation space be $\mathbb{R}^d$ with $d \geq 2$. For any mask operator m that leaves at least one patch visible, there exists an encoder-decoder pair (f, g) that achieves perfect masked reconstruction while all learned representations lie in a one-dimensional subspace of $\mathbb{R}^d$.

To see this, let $v \in \mathbb{R}^d$ be a fixed non-zero vector and let $Q_1, \ldots, Q_N$ be distinct scalars. Since any non-empty visible view $x_i[m]$ contains the patch p_i, and $p_i \neq p_j$ for $i \neq j$, the



(a) singular values (b) effective rank

Figure 2: (a) Comparison of the singular values of learned features with 1) random initialization, 2) MPT loss, and 3) our U-MPT loss.
(b) The changing process of effective rank (Roy and Vetterli, 2007) of the encoded features trained with different objectives (MPT and U-MPT).

visible view uniquely identifies the original image. We can therefore construct an encoder such that

$$f(x_i[m]) = Q_i v.$$

A sufficiently expressive decoder can then map each distinct code $Q_i v$ back to the corresponding image:

$$g(f(x_i[m])) = x_i, \qquad i = 1, \ldots, N.$$

Thus, the reconstruction loss achieves its minimum of zero. However, all representations lie in the one-dimensional subspace span(v). Specifically, the representation matrix satisfies

$$Z = \begin{bmatrix} Q_1 \\ \vdots \\ Q_N \end{bmatrix} v^\top,$$

and hence rank$(Z) = 1$, although the representation space itself is d-dimensional. This counterexample therefore formally demonstrates that optimal masked reconstruction does not, by itself, prevent dimensional collapse.

We further investigate whether such dimensional feature collapse occurs in real-world data by conducting experiments on ImageNet-100. Figure 2(a) shows that after learning, MPT's features exhibit a higher degree of collapse, as indicated by the reduced number of large singular values (representing non-collapse dimensions). Quantitatively, Figure 2(b) shows that during the training process, the features of MPT progressively collapse, resulting in a smaller effective rank (Roy and Vetterli, 2007). This trend confirms that MPT experiences an increasing level of dimensional feature collapse.

5.2 U-MPT: Uniformity-enhanced Masked Pretraining Objective

To further enhance the feature diversity of MPT and alleviate the dimensional collapse issue, we draw inspiration from the uniformity loss in contrastive learning (Chen et al., 2020a; HaoChen et al., 2021) and propose the Uniformity-enhanced MPT (U-MPT) loss, which incorporates an explicit regularization term to promote feature uniformity through a coefficient $\lambda > 0$,

$$\mathcal{L}_{\text{U-MPT}}(h) = \mathcal{L}_{\text{MPT}}(h) + \lambda \cdot \mathcal{L}_{\text{unif}}(f), \tag{15}$$

$$\text{where } \mathcal{L}_{\text{unif}}(f) = \mathbb{E}_{x_1}\mathbb{E}_{x_1^-}(f(x_1)^\top f(x_1^-))^2, \tag{16}$$

and x_1^- denotes an independently drawn unmasked view from $\mathcal{X}_1$. Intuitively, the spectral uniformity loss $\mathcal{L}_{\text{unif}}(f)$ (HaoChen et al., 2021) encourages a small feature similarity between random unmasked views, which could effectively promote the feature diversity of all samples. As a preview of the results, we can observe from Figures 2(a) & 2(b) that U-MPT effectively addresses the dimensional feature collapse issue and significantly improves the effective feature dimensionality by a large margin.

Theoretically, combined with Corollary 5, we can show that the U-MPT loss is lower-bounded by the spectral contrastive loss with a specific choice of λ.

Theorem 8. *Denote the spectral contrastive loss (SCL) from HaoChen et al. (2021) as*

$$\mathcal{L}_{\text{SCL}}(f) = 2\mathcal{L}_{\text{align}}(f) + \mathcal{L}_{\text{unif}}(f) = -2\mathbb{E}_{x_1,x_1^+} f(x_1)^\top f(x_1^+) + \mathbb{E}_{x_1,x_1^-}(f(x_1)^\top f(x_1^-))^2. \tag{17}$$

Under Assumption 1 and the assumption that the decoder is L-bi-Lipschitz, when $\lambda = 1/(4L)$*, the U-MPT loss can be lower bounded by the SCL loss:*

$$\mathcal{L}_{\text{U-MPT}}(h) \geq \frac{1}{4L} \cdot \mathcal{L}_{\text{SCL}}(f) - \varepsilon(g) + \frac{1}{2}. \tag{18}$$

As a result, minimizing the U-MPT loss will implicitly minimize the spectral contrastive loss among input views. This reveals that, by adding a uniformity regularization term, U-MPT aligns well with contrastive learning. Next, we will present the main empirical results of our proposed U-MPT loss on different real-world data sets with different backbones, and then conduct a series of experiments to understand the advantages of U-MPT loss.

5.3 Evaluation on Benchmark Data Sets

To evaluate the effectiveness of the proposed U-MPT loss, extensive experiments are conducted on CIFAR-10 (Krizhevsky and Hinton, 2009), ImageNet-100 (Deng et al., 2009), and ImageNet-1K (Deng et al., 2009) with MAE (He et al., 2022) being the instantiated MPT paradigm.

Setup. We mainly follow the basic setup of MAE: for the encoder, we adopt different variants of ViT (Dosovitskiy et al., 2021), *i.e.,* ViT-Tiny, ViT-Base, and ViT-Large. For the decoder, we use a flexible one following He et al. (2022). The mask ratio is set to 0.75. The coefficient of the uniformity term in the U-MPT loss is set to 0.01. On CIFAR-10, we pretrain the model for 2000 epochs with batch size 4096 and weight decay 0.05. On

ImageNet-100 and ImageNet-1K, we pretrain the model for 200 epochs with batch size 1024 and weight decay 0.05.

In-domain Linear Probing and Fine-tuning. We conduct both linear probing and end-to-end fine-tuning on the pretrained encoder, where pretraining and downstream evaluation use the same data set. For linear probing, the pretrained encoder is frozen and only a linear classifier is optimized, which directly measures the quality of the pretrained representation. As for end-to-end fine-tuning, both the pretrained encoder and the classification head are jointly optimized, which evaluates whether the advantage induced during pretraining persists after full supervised adaptation on the same data distribution. In Table 1, we compare the performance of original MAE loss and U-MPT loss on different benchmarks. We find that, on linear evaluation results, our proposed U-MPT loss average increases 8.9% on CIFAR-10, 7.35% on ImageNet-100, and 3.35% on ImageNet-1K with two different backbones. On fine-tuning results, our proposed U-MPT loss will not hurt the performance of fine-tuning results of MAE. This suggests that full downstream adaptation can reduce the performance gap induced by different pretraining objectives, whereas the advantage of U-MPT is more directly reflected in the quality of the pretrained representations themselves.

		CIFAR-10		ImageNet-100		ImageNet-1K	
Downstream Task	loss	ViT-Tiny	ViT-Base	ViT-Base	ViT-Large	ViT-Base	ViT-Large
Linear Probing	MAE	59.6	61.7	61.2	64.4	55.4	62.2
	U-MPT	**68.9**	**70.2**	**67.5**	**72.8**	**58.5**	**65.8**
Fine-tuning	MAE	**89.6**	90.7	**86.9**	**87.3**	82.9	**83.3**
	U-MPT	89.4	**90.8**	86.8	**87.3**	**83.0**	83.2

Table 1: In-domain linear-probing accuracy (%) and fine-tuning accuracy (%) of models pretrained by MAE loss and U-MPT loss with different ViT backbones on CIFAR-10, ImageNet-100, and ImageNet-1K. The uniformity regularizer term in the U-MPT loss significantly improves the linear evaluation performance of the MAE loss without hurting the performance of fine-tuning.

Cross-dataset Fine-tuning. We next evaluate whether the benefit of U-MPT persists when the pretrained encoder is transferred to and fully adapted on a downstream data set different from the pretraining data set. Starting from models pretrained on ImageNet, we perform end-to-end fine-tuning on four downstream benchmarks with diverse characteristics: 1) iNaturalist-2021 (Van Horn et al., 2021), a large-scale and long-tailed fine-grained species classification data set; 2) CUB (Wah et al., 2011), a fine-grained bird classification benchmark; 3) Cars (Krause et al., 2013), a fine-grained car model classification benchmark; and 4) ImageNet-A (Hendrycks et al., 2021), a challenging data set of naturally occurring images that exhibits a substantial distribution shift from the standard ImageNet distribution. As shown in Table 2, U-MPT improves performance across all four downstream data sets. The improvement is particularly pronounced on iNaturalist-2021, where accuracy increases from 66.0% to 69.4%. On ImageNet-A, U-MPT also improves accuracy from 18.4% to 19.9%. These results suggest that the benefit of U-MPT can remain visible even after the pretrained encoder is fully adapted to a shifted downstream data set.

Method	iNat_{21}	CUB	Cars	ImageNet-A
MAE	66.0	81.5	59.8	18.4
U-MPT	**69.4**	**81.8**	**60.3**	**19.9**

Table 2: Cross-dataset fine-tuning accuracies (%) of models pretrained on ImageNet using MAE and U-MPT objectives. The pretrained models are subsequently fine-tuned and evaluated on different downstream data sets.

Data Set and Corruption Type	MAE	U-MPT
ImageNet-A	0.57	**0.68**
ImageNet-C (Defocus Blur)	10.48	**26.20**
ImageNet-C (Motion Blur)	16.36	**35.84**
ImageNet-C (Frosted Glass Blur)	13.68	**30.14**
ImageNet-C (Zoom Blur)	14.24	**32.48**
ImageNet-C (Elastic Transform)	23.02	**39.98**
ImageNet-C (Contrast)	9.76	**23.40**
ImageNet-C (Brightness)	28.52	**50.14**
ImageNet-C (Pixelate)	24.66	**46.18**
ImageNet-C (Snow)	9.86	**22.84**
ImageNet-C (Fog)	10.52	**24.62**
ImageNet-C (Frost)	11.30	**23.64**
ImageNet-C (JPEG Compression)	20.36	**39.24**
ImageNet-C (Gaussian Noise)	11.46	**24.94**
ImageNet-C (Impulse Noise)	8.78	**20.08**
ImageNet-C (Shot Noise)	11.40	**24.08**

Table 3: Out-of-distribution generalization accuracies (%) of MAE and U-MPT on ImageNet-A and ImageNet-C. The pretrained encoder is kept frozen and evaluated on ImageNet-A and ImageNet-C using linear probing.

Out-of-Distribution Generalization. We further evaluate the quality of the frozen ImageNet pretrained representations under distribution shift using ImageNet-A (Hendrycks et al., 2021) and ImageNet-C (Hendrycks and Dietterich, 2019). Specifically, models are pretrained on ImageNet, after which the pretrained encoder is kept frozen and evaluated on ImageNet-A and ImageNet-C using linear probing. In contrast to the cross-dataset fine-tuning setting above, the encoder is not adapted to the target distribution, allowing us to more directly assess the generalization ability of the pretrained representations under distribution shift. The results presented in Table 3 demonstrate that incorporating the uniformity loss significantly improves the classification accuracies on ImageNet-A and each

corruption type of ImageNet-C. We believe that this improvement is due to the uniformity loss encouraging greater feature diversity, which helps to separate the feature representations of different classes. Consequently, when corruptions are applied, the features are less likely to be misclassified into the feature space of another class. Similar connections between improved feature separation and robustness under distribution shifts have also been observed in prior work (Zhang et al., 2025a).

Extension to Other MIM Frameworks. We also verify our uniformity-enhanced loss in another MIM framework SimMIM (Xie et al., 2022). For SimMIM, we use ViT-Base as the encoder and use the linear decoder as in Xie et al. (2022). We use the recommended mask ratio 0.6. The coefficient of the uniformity term is set to 0.01, the same as in MAE. For ImageNet-100, we pretrain the model for 200 epochs with batch size 128 and weight decay 0.05. Linear evaluation is conducted in Table 4. We can see that the linear accuracy increases by 6.8% with the uniformity regularizer, which further verifies the general property of our approach.

SimMIM	With Uniformity Loss
54.3	**61.1**

Table 4: Linear probing accuracy (%) of uniformity-enhanced SimMIM on ImageNet-100.

λ	0	1e-3	1e-2	1e-1	1e0
Acc	61.2	65.9	**67.5**	45.1	47.0

Table 5: Linear probing accuracy (%) of uniformity-enhanced MAE with different coefficients.

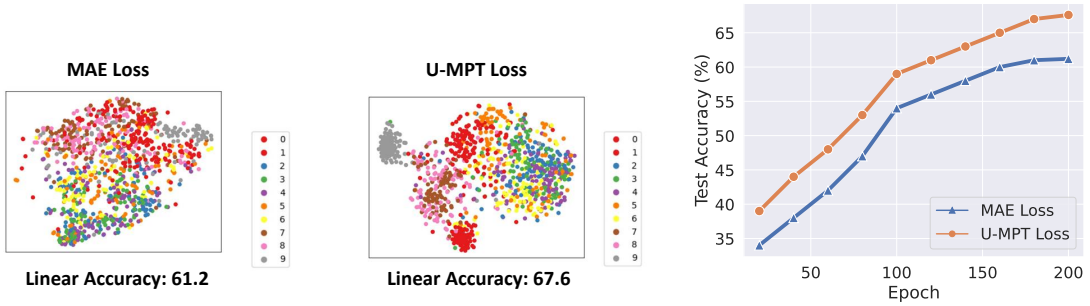
5.4 Empirical Understandings

To further understand the empirical behavior of U-MPT, we analyze the effect of the regularization coefficient, the geometry of the learned representations, and the training dynamics.

Different Coefficients of the Regularizer Term. The most important hyper-parameter of our proposed U-MPT loss is the coefficient of the uniformity regularizer term. In Table 5, we present the results of linear evaluation on ImageNet-100 trained with the U-MPT loss with different coefficients of the regularizer term. We can see that the downstream performance increases when the coefficient increases from 0 to 0.01. However, the overlarge coefficient will also hurt the performance of U-MPT loss as the task of MAE will be overlooked.

Visualization of Representations. To intuitively understand the improvement of our U-MPT loss on clustering intra-class samples, we use t-SNE (van der Maaten and Hinton, 2008) to visualize the representations trained with MAE loss and U-MPT loss on ten random classes of ImageNet-100. As shown in Figure 3(a), we find that with our uniformity regularizer term, the samples are much better-clustered corresponding to their ground-truth labels. To be specific, the red class ("hens") and the gray class ("indigo birds") are separated from others, this is because most of other classes are the animals living in the oceans while these two classes are more like the birds living on the land or the sky. Thus, these two classes are easier to distinguish, especially with our uniformity regularizer.

Training Curve. To further compare the performance between the original MAE loss and U-MPT loss, we plot the linear evaluation accuracy on ImageNet-100 during the training process in Figure 3(b). We can observe that our proposed U-MPT loss improves the



(a) t-SNE representations

(b) Comparison between original MAE loss and our U-MPT loss along training.

Figure 3: (a) Visualization of representations on random 10 classes of ImageNet-100 trained with MAE loss and our U-MPT loss. Our U-MPT loss significantly improves the class-clustering performance of the encoder. (b) The linear evaluation results during the training process. It shows that our U-MPT loss improves the downstream performance consistently along training.

performance of MAE with all different training epochs, which verifies the effectiveness of our proposed U-MPT loss.

6. Further Theoretical Generalization Analysis

Having established that MPT implicitly aligns positive pairs through its masking mechanism, we now turn to a theoretical analysis of its downstream generalization performance based on this connection. Without loss of generality, we take the pixel-wise U-MPT objective as an example. In Section 6.1, we derive the first theoretical guarantees linking MPT's pretraining loss to downstream classification error. In Section 6.2, we reveal how the mask ratio ρ influences the downstream generalization error through a toy model. Furthermore, based on the theoretical analysis, in Section 6.3, we propose a surrogate metric to estimate the downstream error and find that it correlates well with the empirical designs of MPT.

6.1 Theoretical Guarantees on Downstream Classification

In this part, we analyze the downstream performance of U-MPT on the c-class linear classification task (Saunshi et al., 2019; HaoChen et al., 2021). The performance is measured by the prediction accuracy of natural images $\bar{x}$ *w.r.t.* their labels $y(\bar{x})$ when applying a linear prediction head upon pretrained features. For simplicity, we adopt the mean classifier $p_f(x) = \arg\max W_f f(x)$, where $W_f \in \mathbb{R}^{c\times k}$ is the weight of the linear classification head, and for $y \in [c]$, the y-th row $W_y = \mathbb{E}_{x_1|y} f(x_1)^\top$ contains the mean representation of the class y. It has been empirically demonstrated by Saunshi et al. (2019) that the mean classifier achieves comparable performance to learnable linear heads. And we define an ensembled linear predictor p'_f. For an original sample $\bar{x}$, the predictor ensembles the predictions from

all different views and selects the label with the highest frequency. By default, we set the uniformity coefficient $\lambda = 1/(4L)$ as in Theorem 8.

Theorem 9. *Denote the mask-induced label error as $\alpha = \mathbb{E}_{\bar{x},x_1}\mathbb{1}[y(x_1) \neq y(\bar{x})]$. Then, for $\forall\ h \in \mathcal{H}$ (the hypothesis class) with $h = g \circ f$, the downstream classification error of its encoder can be upper bounded by its U-MPT pretraining loss:*

$$\Pr\left(y(\bar{x}) \neq p'_f(\bar{x})\right) \leq c_1 L \cdot \mathcal{L}_{\text{U-MPT}}(h) + c_2\alpha + c_3 L\varepsilon(g) + c_4, \tag{19}$$

where $c_1, \ldots, c_4$ are constants and $c_3 > 1$.

This theorem establishes an upper bound on the downstream error of an encoder f trained with the U-MPT pretraining loss, which is the *first theoretical guarantee* on the downstream performance of MPT methods. As an implication of this theorem, a small U-MPT loss would provably imply a small downstream classification error, which helps explain the strong downstream generalization ability of MAE (He et al., 2022). Additionally, we establish a common lower bound on the U-MPT loss that holds for all $h \in \mathcal{H}$. As a large common lower bound of U-MPT loss indicates that the downstream error will always have a large upper bound, we should pursue a small common lower bound in the following theorem.

Theorem 10. *The U-MPT pretraining loss has the following common lower bound:*

$$\forall\ h \in \mathcal{H}, \quad \mathcal{L}_{\text{U-MPT}}(h) \geq \frac{1}{4L}\sum_{i=k+1}^{N_1} \lambda_i^2 - \varepsilon(g) + \text{const}, \tag{20}$$

where $\lambda_1 \geq \cdots \geq \lambda_{N_1}$ denote the eigenvalues of A.

Combining Theorem 9 and Theorem 10, it is evident that the downstream error of MPT can be minimized with three factors: a small vanilla autoencoding error $\varepsilon(g)$, a small label error α, and smaller magnitude of "residual eigenvalues", *i.e.,* $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$. Here, residual eigenvalues represent the high-frequency components of the augmentation graph $\mathcal{G}_A$ that cannot be fitted by k-dimensional features. The vanilla autoencoding error $\varepsilon(g)$ depends on the non-degeneracy of the decoder g, while the label error α and residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$ are purely reliant on the augmentation graph $\mathcal{G}_A$ induced by the applied mask. Consequently, overall speaking, a capacious MPT model can have reduced downstream errors with a diminished vanilla autoencoding error $\varepsilon(g)$. In the meantime, the masking ratio ρ should be properly chosen to ensure a small label error α and small magnitude of residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$. In fact, as demonstrated by He et al. (2022), the masking ratio exerts a pivotal sway over the downstream efficacy of MAE. Therefore, in the next part, we explore how the choice of mask ratio would affect the downstream generalization of MPT via the label error α and residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$.

6.2 Theoretical Analysis on the Effect of Mask Ratio

To characterize the effects of the mask ratio ρ on label error and graph connectivity, we analyze a binary classification task on a spatially structured simulated data set. As illustrated in Figure 4, each image consists of K patch positions arranged on a two-dimensional grid. For class C_i and position k, the patch x_k is sampled from a position-specific candidate set

$P_{i,k} = E_{i,k} \cup O_k$, where $E_{i,k}$ contains N_k class-specific patches and O_k contains M_k patches shared by the two classes. The candidate identities vary across positions, thereby incorporating spatial locality into the image construction. To deliver a mathematically tractable model, we further assume $N_k = N$ and $M_k = M$ for each k. During masked pretraining, for a mask m with mask ratio ρ, we select ρK positions to form the masked view, while the remaining $(1-\rho)K$ positions constitute the unmasked view. Using this formulation, we explicitly construct the mask-induced augmentation graph and analyze how ρ influences the label error α, the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{N_1}$, and the generalization bound.

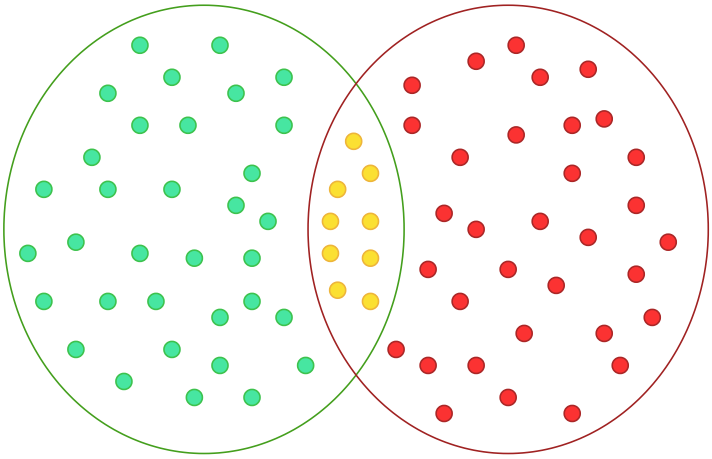


Figure 4: An illustration of the toy model. Each dot represents a patch in real-world data. The green and red dots correspond to non-overlapped patches from different classes, while the yellow dots are overlapped patches.

Theorem 11. *Let $r = \frac{M}{N+M}$ be the ratio of shared patches in each class, and ρ the mask ratio satisfying $0 < \rho < 1$. When $r^K < \frac{1}{12}$ and $0.5 \le \rho \le 1$, with a larger mask ratio, the label error (α) increases monotonically, the residual eigenvalues ($\sum \lambda_i^2$) decrease monotonically, and the downstream error bound decreases first and then increases. Besides, the upper bound for $\Pr(y(\bar{x}) \neq p'_f(\bar{x}))$ achieves its minimum value when $0.5 < \rho < 1$.*

As shown in Theorem 11, the label error increases as the mask ratio ρ increases. Meanwhile, the sum of squared residual eigenvalues exhibits a symmetric trend around $\rho = 0.5$ and decreases as ρ increases beyond 0.5. Consequently, there exists a trade-off between the label error and the eigenvalues when $\rho > 0.5$. Theorem 11 further proves that the generalization bound first decreases and then increases with the mask ratio, and the optimal point lies in the intermediate range. This finding aligns with empirical results from MAE pretraining (He et al., 2022), i.e., 75% is the best mask ratio for downstream tasks. Notably, while there is a local minimum in the generalization bound for $\rho < 0.5$, this point does not achieve a lower bound than the optimal point for $\rho > 0.5$.

Theoretical Insights. As discussed above, Theorems 9 & 10 show the significance of having a small label error α and small residual eigenvalues $\lambda_{k+1}, \ldots, \lambda_{N_1}$ for good downstream MPT performance. Furthermore, Theorem 11 demonstrates that the mask ratio ρ has a decisive influence on both factors. On one hand, the label error α tends to increase with a larger mask ratio. Intuitively, α indicates the likelihood of accurately recovering the original class y from the masked portions. Figure 5 illustrates that small or moderate mask ratios, even a substantial one like 0.75, barely alter the class identity. Only when the mask ratio is exceptionally high, for instance, 0.95, the object's identity becomes nearly unrecognizable, such as with a car in the example. On the other hand, according to spectral graph theory (Chung, 1997), the residual eigenvalues $\lambda_{k+1}, \ldots, \lambda_{N_1}$ reflect the high-frequency aspects of the graph, possessing significant magnitudes when the graph is less interconnected (with numerous disjoint components). Consequently, a low downstream error necessitates enhanced connectivity within the augmentation graph $\mathcal{G}_A$. Figure 5 visually demonstrates that elevating the mask ratio masks out many diverse patterns, thereby augmenting similarity among remaining patches, especially within intra-class samples (e.g., the tires of two cars). Hence, a higher mask ratio effectively improves graph connectivity by diminishing sample diversity and augmenting inter-sample resemblance. This observation is also consistent with Zhang et al. (2024), which shows that the flexible prediction targets in masked pretraining can induce richer inter-sample connections. Considering the impacts on both the label error α and the residual components $\lambda_{k+1}, \ldots, \lambda_{N_1}$, Theorem 11 shows an evident trade-off in selecting the mask ratio: we should choose a mask ratio that is appropriately large to enhance graph connectivity. However, we also need to avoid excessively large mask ratios that lead to class mixture and confusion. In other words, a guiding principle emerges: the mask ratio should be chosen such that mask-induced graph connectivity primarily occurs among *intra-class* samples (no harms), rather than *inter-class* samples (resulting in significant label errors).

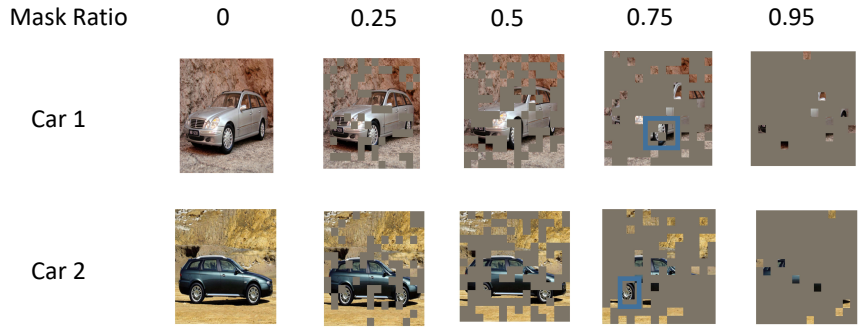


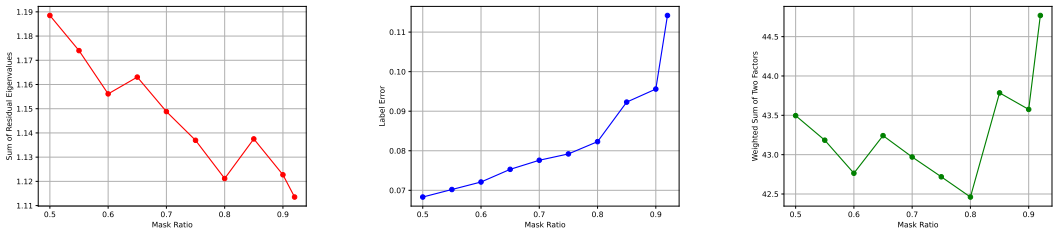
Figure 5: An appropriate mask ratio can generate similar views from different samples in the same class.

6.3 Empirical Estimation of Downstream Errors

To further validate our argument, we introduce a metric to empirically estimate the downstream error of MPT methods on real-world data sets. Since real-world images are sampled from a continuous space and two patches cannot be identical, we cannot directly estimate

the probability that unmasked views share the overlapped masked view. Consequently, we first train a Vector Quantised-Variational AutoEncoder (VQ-VAE) model (van den Oord et al., 2017) and map patches into discrete tokens. Subsequently, we randomly selected 2000 images from the same data set and applied 5 random masks to each, resulting in a total of 10000 nodes within the augmentation graph. For all unmasked regions, we utilized the VQ-VAE model to map each visible patch to a unique patch ID, thus representing each view as a multiset of these IDs.

The edge weights between any two unmasked views were then estimated by calculating the degree of overlap between the sets of patch IDs from their respective masked counterparts. This degree of overlap was defined as the size of the intersection of patch IDs divided by the total number of patches in a masked view. Graph connectivity was quantified by the sum of squared residual eigenvalues, specifically $\frac{1}{N_1}\sum_{i=k+1}^{N_1}\lambda_i^2$, where k represents the feature dimension (set to 1000 in this case based on common practice) and N_1 denotes the number of nodes in the augmentation graph (equal to 10000). The normalization coefficient $\frac{1}{N_1}$ is introduced to mitigate the impact of the number of nodes selected. Additionally, the label error was estimated by the proportion of unmasked-view pairs that exhibited over 90% patch overlap yet possessed different ground-truth labels. The empirical results are illustrated in Figure 6.



(a) Residual eigenvalues (b) Label Error (c) Tradeoff

Figure 6: (a)(b) Estimated residual eigenvalues of the augmentation graph and label error for 2000 randomly selected images from the ImageNet-100 data set, with each image subjected to 5 different random masks, under varying mask ratios.
(c) The blended value of these two factors.

The results demonstrate a steady decline in the residual eigenvalues as the mask ratio increases from 0.5, which indicates that a higher mask ratio effectively improves the connectivity of the augmentation graph. Meanwhile, the label error shows a general upward trend, which is a result of growing vagueness as more and more information is masked out. To quantify their respective contributions, we adopt the coefficients given in Section A.5, *i.e.*, $\frac{32}{N_1}\sum_{i=k+1}^{N_1}\lambda_i^2 + 80\alpha$, and find that the weighted sum reaches its smallest value when the mask ratio is around 0.8, very close to the value employed in He et al. (2022). These observations align with our theoretical analysis, highlighting the inherent trade-off between the graph connectivity and the label error.

7. Theory Inspired Strategies for Performance Improvement of MPT

As discussed in the theoretical framework, we introduce a principle to improve the performance of MPT, i.e., we should enhance connectivity between unmasked views while avoiding label errors. Based on that, we propose new theory-guided strategies to improve the performance of MPT and analyze current variants of MPT from this perspective. In Section 7.1, we present a novel approach that involves reassigning weights for each patch in the reconstruction loss to enhance the connectivity of the augmentation graph. In Section 7.2, we review existing strategies with our theoretical principles.

7.1 Patch Re-weighting

In original MPT approaches, the weight for each masked patch in the reconstruction loss is uniform. However, as analyzed in contrastive learning, there exist intra-class samples that are difficult to be pulled together in the feature space and these samples usually have a greater impact on forming the connectivity of the augmentation graph (Zhang et al., 2026). Inspired by this, we consider paying more attention to the patches that are hard to reconstruct in MPT, namely the patches that are difficult to implicitly align.

Specifically, we propose reweighted MAE, where the weight for a patch in the reconstruction loss is dynamically adjusted based on the reconstruction loss of that specific patch. Let h denote the autoencoder, x_1 the unmasked view, and x_2 the masked view. Here, $h(x_1)$ represents the reconstructed masked view given x_1, while $h(x_1)_k$ and $x_{2,k}$ denote the k-th patch of $h(x_1)$ and x_2, respectively. The reconstruction loss for the k-th patch is defined as:

$$l_k(x_1, x_2) = \|h(x_1)_k - x_{2,k}\|_2^2.$$

The original MAE reconstruction loss for x_1 and x_2 is then formulated as:

$$\mathcal{L}_{\text{MAE}}(h, x_1, x_2) = \sum_k l_k(x_1, x_2).$$

To introduce patch re-weighting, we define the weight for the k-th patch as:

$$w_k = [\text{stopgrad}(l_k(x_1, x_2))]^\gamma,$$

where $\text{stopgrad}(\cdot)$ denotes the stop gradient operation, ensuring that w_k is detached from the back-propagation process. The reweighted loss function is then given by:

$$\mathcal{L}_{\text{Re-weighted}} = \frac{1}{\sum_k w_k} \sum_k w_k \cdot l_k(x_1, x_2).$$

To evaluate the effectiveness of re-weighted MAE, we conduct a series of experiments using a ViT-Base encoder pretrained on ImageNet-1K. As shown in Table 6, re-weighted MAE with $\gamma = 1.5$ improves the in-domain linear probing accuracy on ImageNet-1K from 55.4% to 58.6%. The improvement is more pronounced under distribution shifts, where the mean accuracy across the fifteen ImageNet-C corruption types increases from 14.96% to 22.24%. These results suggest that emphasizing harder-to-reconstruct patches can improve both the quality and robustness of the learned representations.

Setting	Data Set	Vanilla MAE	Re-weighted MAE
In-domain	ImageNet-1K	55.4	**58.6**
OOD	ImageNet-A	0.57	**0.87**
	ImageNet-C (Defocus Blur)	10.48	**20.51**
	ImageNet-C (Motion Blur)	16.36	**24.28**
	ImageNet-C (Frosted Glass Blur)	13.68	**19.03**
	ImageNet-C (Zoom Blur)	14.24	**19.21**
	ImageNet-C (Elastic Transform)	23.02	**30.94**
	ImageNet-C (Contrast)	9.76	**20.41**
	ImageNet-C (Brightness)	28.52	**46.42**
	ImageNet-C (Pixelate)	24.66	**25.50**
	ImageNet-C (Snow)	9.86	**16.93**
	ImageNet-C (Fog)	10.52	**20.69**
	ImageNet-C (Frost)	11.30	**22.00**
	ImageNet-C (JPEG Compression)	20.36	**29.41**
	ImageNet-C (Gaussian Noise)	11.46	**14.82**
	ImageNet-C (Impulse Noise)	8.78	**9.81**
	ImageNet-C (Shot Noise)	11.40	**13.60**
	Mean over ImageNet-C	14.96	**22.24**

Table 6: Results (%) of vanilla MAE and Re-weighted MAE with $\gamma = 1.5$, where OOD means out-of-distribution. Both models are pretrained on ImageNet-1K, and the learned representations are evaluated via linear probing on ImageNet-1K for in-domain performance and on ImageNet-A and ImageNet-C for out-of-distribution performance.

We further conduct an ablation study to investigate the effect of the re-weighting coefficient γ on ImageNet-100. As shown in Figure 7, the linear probing accuracy first increases as γ grows and then decreases when γ becomes too large, with the best performance achieved at $\gamma = 1.5$. Specifically, $\gamma = 1.5$ improves the linear probing accuracy by 1.98% over vanilla MAE (i.e., $\gamma = 0$). This trend suggests that moderately emphasizing harder-to-reconstruct patches is beneficial, whereas overly large weights may overemphasize a small subset of difficult patches and hurt the overall representation quality.

7.2 Explaining Existing Strategies

Recent improvements to MPT broadly fall into two categories: refinements to masking strategies and integration with contrastive learning. We find that most of these improvements can be explained with our theoretical framework (Section 6), which emphasizes the roles of label error (α) and augmentation graph connectivity ($\sum \lambda_{k+1}^2$).

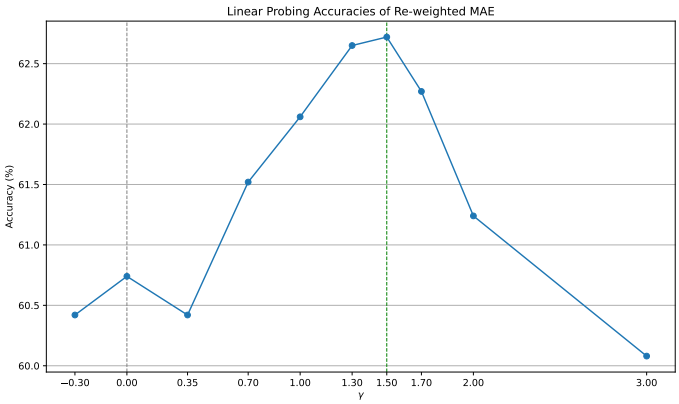


Figure 7: Linear probing accuracies of ViT-Base pretrained using Re-weighted MAE on the ImageNet-100 data set for 200 epochs.

Masking Strategy Improvements. A prominent line of work optimizes masking by prioritizing *salient patches*—regions critical for downstream tasks. For instance, Liu et al. (2023) proposed guided masking using attention scores from the CLS token in Vision Transformers. Periodically, unmasked images are passed through the encoder to compute an importance map, where higher attention scores indicate salient patches (e.g., foreground objects). These patches are then masked with higher probability, while less important regions (e.g., backgrounds) are retained as encoder inputs. Wang et al. (2023a) dynamically adjust masking based on the reconstruction difficulty of the patches and preferentially mask the core objects. Expanding this concept to relational learning, Yang et al. (2023) introduced masked relation modeling for medical images. This approach leverages the self-attention mechanism to detect highly interdependent regions within the input image and disrupts these relationships by masking the most critical patches associated with a specific region. These methods align well with our theoretical framework. By masking salient patches, they reduce the distinctness of masked views, generating more intra-class edges in the augmentation graph $\mathcal{G}_A$. This lowers residual eigenvalues ($\sum \lambda_{k+1}^2$) and improves downstream performance as Theorem 10.

Contrastive Learning Integration. Another category of improving strategy is to combine MPT with contrastive objectives. Huang et al. (2024) proposed CMAE, which exemplifies this through a dual-branch architecture. The first branch follows the standard MAE framework: an encoder processes masked and pixel-shifted images, and a lightweight decoder reconstructs pixels. The second branch introduces a momentum encoder updated via exponential moving average (EMA), which processes the unshifted and unmasked version of inputs. A contrastive loss aligns features from the EMA encoder with those decoded by the first branch, effectively distilling knowledge between masked and unmasked views. The authors report a 5.9% improvement in ImageNet-1K linear probing accuracy over MAE, attributing this to reduced intra-class variance and enhanced inter-class diversity in learned features. Jiang et al. (2023) proposed Layer-Grafted Pretraining for Vision Transformers. Early layers are trained with MIM to capture low-level textures, while later layers employ

contrastive learning to enhance semantic discriminability. This is achieved iteratively: after pretraining early layers with reconstruction loss, they are frozen, and contrastive objectives are applied to later layers. The authors observed improved separability between different classes following the application of Layer-Grafted Pretraining, based on a qualitative evaluation of the t-SNE visualizations of the representations. We note that MPT generates intra-class edges by masking techniques while contrastive learning generates edges by data augmentation. Consequently, combining these two objectives can generate more edges and further enhance the connectivity of the augmentation graph, which indicates better downstream performance.

8. Conclusion

In this work, we established a theoretical connection between masked pretraining (MPT) and contrastive learning, showing that minimizing the MPT loss implicitly aligns mask-induced positive pairs. We identified dimensional collapse as a key limitation of MPT and proposed U-MPT, a uniformity-enhanced variant that explicitly promotes feature diversity. Theoretically, we derived downstream generalization guarantees, explaining the empirical success of high mask ratios (e.g., $\rho = 0.75$) through their balance of intra-class connectivity and inter-class discriminability. Empirically, U-MPT achieved superior robustness on out-of-distribution data sets like ImageNet-A and ImageNet-C. Additionally, based on our theoretical analysis, we propose a new reconstruction strategy that prioritizes challenging regions during reconstruction and enhances the performance of MPT. These contributions provide a unified framework for optimizing MPT, bridging reconstruction fidelity with feature discriminability.

Acknowledgments

Yisen Wang was supported by National Natural Science Foundation of China (92370129, 62376010), Beijing Major Science and Technology Project under Contract no. Z251100008425006, Beijing Natural Science Foundation (L257007), Beijing Nova Program (20230484344, 20240484642), and State Key Laboratory of General Artificial Intelligence.

Appendix A. Proofs

In the following proofs, given the fact that g is fixed, we write $\varepsilon = \varepsilon(g)$ for notational simplicity.

A.1 Proof of Theorem 4

Proof With Assumption 1, we have

$$\begin{aligned}
\mathcal{L}_{\mathrm{MAE}}(h) &= \mathbb{E}_{x_1,x_2}\|h(x_1) - x_2\|^2 \\
&= \mathbb{E}_{x_1,x_2}\|h(x_1) - x_2\|^2 + \varepsilon - \varepsilon \\
&\geq \mathbb{E}_{x_1,x_2}\|h(x_1) - x_2\|^2 + \|x_2 - h_g(x_2)\|^2 - \varepsilon && (\|x_2 - h_g(x_2)\|^2 \leq \varepsilon) \\
&\geq \frac{1}{2}\mathbb{E}_{x_1,x_2}\|h(x_1) - h_g(x_2)\|^2 - \varepsilon && (\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)) \\
&= -\mathbb{E}_{x_1,x_2} h(x_1)^\top h_g(x_2) - \varepsilon + 1. && (h(x) \text{ and } h_g(x) \text{ are normalized})
\end{aligned}$$

We formulate the features as two matrices H,H_g. We denote $H(x_1) = \sqrt{d_{x_1}}h(x_1)$ as the x_1-th row of the matrix H and $H_g(x_2) = \sqrt{d_{x_2}}h_g(x_2)$ as the x_2-th row of the matrix H_g. As defined before, $(A_M)_{x_2,x_1}$ is the joint distribution of x_1 and x_2, *i.e.,* $(A_M)_{x_2,x_1} = w_{x_1,x_2}$. We denote the normalized form of A_M as $\bar{A}_M$, *i.e.,* $(\bar{A}_M)_{x_2,x_1} = \frac{w_{x_1,x_2}}{\sqrt{d_{x_1}}\sqrt{d_{x_2}}}$. Then we can reformulate the reconstruction loss,

$$\begin{aligned}
\mathcal{L}_{\mathrm{MAE}}(h) &\geq -\sum_{x_1,x_2} \frac{w_{x_1,x_2}}{\sqrt{d_{x_1}d_{x_2}}}\sqrt{d_{x_1}}h(x_1) \cdot \sqrt{d_{x_2}}h_g(x_2) - \varepsilon + 1 \\
&= -\operatorname{tr}(\bar{A}_M H H_g^\top) - \varepsilon + 1 \\
&\geq -\frac{1}{2}(\|\bar{A}_M H\|^2 + \|H_g\|^2) - \varepsilon + 1.
\end{aligned}$$

As the output of decoder is normalized, *i.e.,* $\|H_g\|^2 = 1$. we obtain

$$\mathcal{L}_{\mathrm{MAE}}(h) \geq -\frac{1}{2}(\operatorname{tr}(\bar{A}_M^\top \bar{A}_M H H^\top) + 1) - \varepsilon + 1 = -\frac{1}{2}\operatorname{tr}(\bar{A}_M^\top \bar{A}_M H H^\top) - \varepsilon + \frac{1}{2}. \tag{21}$$

Then we element-wise compute $\bar{A}_M^\top \bar{A}_M$ and $HH^\top$, we have

$$(\bar{A}_M^\top \bar{A}_M)_{x_1,x_1^+} = \sum_{x_2} \frac{w_{x_1,x_2} w_{x_1^+,x_2}}{d_{x_2}\sqrt{d_{x_1}d_{x_1^+}}}. \tag{22}$$

$$(HH^\top)_{x_1^+,x_1} = \sqrt{d_{x_1}d_{\ x_1^+}}h(x_1)^\top h(x_1^+). \tag{23}$$

As the trace is the sum of the diagonal value of the matrix, we consider x_1-th diagonal value of $(\bar{A}_M^\top \bar{A}_M H H^\top)$, *i.e.,*

$$(\bar{A}_M^\top \bar{A}_M H H^\top)_{x_1,x_1} = \sum_{x_1^+}(A_M^\top A_M)_{x_1,x_1^+}(HH^\top)_{x_1^+,x_1} = \sum_{x_1^+}\sum_{x_2} \frac{w_{x_1,x_2} w_{x_1^+,x_2}}{d_{x_2}} h(x_1)^\top h(x_1^+). \tag{24}$$

With that, we can element-wise expand Eq. (21),

$$\mathcal{L}_{\text{MAE}}(h) \geq -\frac{1}{2}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x,x^+)}h(x_1)^\top h(x_1^+) - \varepsilon + \frac{1}{2},$$

where $\hat{p}(x_1, x_1^+) = \sum_{x_2} \frac{w_{x_1,x_2} w_{x_1^+,x_2}}{d_{x_2}}$. Similarly, we define a matrix H_g, where $(H_g)_{x_2} = \sqrt{d_{x_2}} h_g(x_2)$. We let $(\bar{A}_M)_{x_2,x_1} = \frac{w_{x_1,x_2}}{\sqrt{d_{x_1}}\sqrt{d_{x_2}}}$ and we obtain

$$\begin{aligned}
\mathcal{L}_{\text{MAE}}(h) &\geq -\operatorname{tr}(H H_g^\top \bar{A}_M) - \varepsilon + 1 \\
&\geq -\frac{1}{2}(\|H\|^2 + \|\bar{A}_M^\top H_g\|^2) - \varepsilon + 1 && (tr(AB) \leq \frac{1}{2}\|A\|^2 + \|B\|^2) \\
&= -\frac{1}{2}(\operatorname{tr}(\bar{A}_M \bar{A}_M^\top H_g H_g^\top) + 1) - \varepsilon + 1 && (H_g \text{ is normalized}) \\
&\geq -\frac{1}{2}\mathbb{E}_{x_2^+\sim\bar{p}(x_2,x_2^+)} h_g(x_2)^\top h_g(x_2^+) - \varepsilon + \frac{1}{2},
\end{aligned}$$

where $\bar{p}(x_2, x_2^+) = \sum_{x_1} \frac{w_{x_1,x_2} w_{x_1,x_2^+}}{d_{x_1}}$. ■

A.2 Proof of Corollary 5

Proof With Theorem 4, we know that

$$\mathcal{L}_{\text{MAE}}(h) \geq -\frac{1}{2}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x,x^+)}h(x_1)^\top h(x_1^+) - \varepsilon + \frac{1}{2}.$$

As the decoder is L-bi-Lipschitz, we obtain

$$\forall\ (x_1, x_2), 1/L \cdot \|x_1 - x_2\|^2 \leq \|g(x_1) - g(x_2)\|^2 \leq L \cdot \|x_1 - x_2\|^2. \tag{25}$$

So,

$$\begin{aligned}
\mathcal{L}_{\text{MAE}}(h) &\geq -\frac{1}{2}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} h(x_1)^\top h(x_1^+) - \varepsilon + \frac{1}{2} \\
&= \frac{1}{4}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} \|h(x_1) - h(x_1^+)\|^2 - \varepsilon && (h(x) \text{ is normalized}) \\
&= \frac{1}{4}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} \|g(f(x_1)) - g(f(x_1^+))\|^2 - \varepsilon \\
&\geq \frac{1}{4L}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} \|f(x_1) - f(x_1^+)\|^2 - \varepsilon && (\text{Equation 25}) \\
&= -\frac{1}{2L}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} f(x_1)^\top f(x_1^+) - \varepsilon + \frac{1}{2}.
\end{aligned}$$

■

A.3 Proof of Theorem 6

Proof When the encoder fully collapses, the encoder f maps all the features to the same point c, *i.e.,*

$$\forall x \in \mathcal{X}_1, f(x) = c. \tag{26}$$

Then,

$$\begin{aligned} \mathcal{L}_{MAE}(h) &= \mathbb{E}_{x_1,x_2}\|g(f(x_1)) - x_2\|^2 \\ &= \mathbb{E}_{x_2}\|g(c) - x_2\|^2. \end{aligned} \tag{27}$$

We then select a g(c) to make Equation (27) minimal. According to KKT conditions, it has a closed-form solution $q^\star$, satisfying

$$2\mathbb{E}_{x_2}(q^\star - x_2) = 0. \tag{28}$$

i.e., $q^\star = \mathbb{E}_{x_2} x_2$. Then

$$\begin{aligned} \mathcal{L}_{MAE}(h) &= \mathbb{E}_{x_2}\|g(c) - x_2\|^2 \\ &\geq \mathbb{E}_{x_2}\|\mathbb{E}_{x_2'} x_2' - x_2\|^2 \\ &= \mathrm{Var}(x_2). \end{aligned}$$

■

A.4 Proof of Theorem 8

Proof With Corollary 5, we have

$$\mathcal{L}_{\mathrm{MAE}}(h) \geq -1/(2L) \cdot \mathbb{E}_{x_1,x_1^+} f(x_1)^\top f(x_1^+) - \varepsilon + \mathrm{const}. \tag{29}$$

Then we set $\lambda = \frac{1}{4L}$ and we obtain

$$\begin{aligned} \mathcal{L}_{\text{U-MPT}}(h) &= \mathcal{L}_{MAE}(h) + 1/(4L) \cdot \mathcal{L}_{unif}(f) \\ &\geq 1/(2L) \cdot \mathcal{L}_{align}(f) + 1/(4L) \cdot \mathcal{L}_{unif}(f) - \varepsilon + \mathrm{const} \\ &= 1/(4L) \cdot \mathcal{L}_{SCL}(f) - \varepsilon + \mathrm{const}. \end{aligned} \tag{30}$$

■

A.5 Proof of Theorem 9

Proof

We compose the marginal distribution of x_1 as a matrix D and $D_{x_1} = d_{x_1}$ is the x_1-th row of D. And we denote U as the matrix composed of encoder features, *i.e.,* $U_{x_1} = \sqrt{d_{x_1}} f(x_1)$. Recall that $A_{x_1,x_1^+} = \mathcal{A}(x_1, x_1^+)$ and $\bar{A}$ is the normalized form of A, *i.e.,* $\bar{A}_{x_1,x_1^+} = \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1} \cdot d_{x_1^+}}}$.

Then we reformulate the downstream error,

$$\begin{aligned}\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 &= \sum_{(x_1, y_{x_1})} d_{x_1}\|y_{x_1} - W_f f(x_1)\|^2 \\ &= \|D^{1/2}Y - UW_f^\top\|^2 \\ &= \|D^{1/2}Y - \bar{A}C + \bar{A}C - UW_f^\top\|^2,\end{aligned}$$

where $C_{x_1,j} = \sqrt{(d_i)\mathbb{1}_{y_{x_1}=j}}$. Then we consider the relationship between the downstream error and the augmentation graph, we element-wise consider the matrix $(D^{1/2}Y - \bar{A}C)$,

$$(D^{1/2}Y)_{x_1,j} = \sqrt{(d_{x_1})\mathbb{1}_{y_{x_1}=j}}, \quad (\bar{A}C)_{x_1,j} = \sum_{x_1^+} \frac{w_{x_1,x_1^+}}{\sqrt{d_{x_1}}\cdot\sqrt{d_{x_1^+}}}\sqrt{(d_{x_1^+})\mathbb{1}_{y_{x_1^+}=j}}. \tag{31}$$

So when $j = y_{x_1}$,

$$\begin{aligned}(D^{1/2}Y - \bar{A}C)_{x_1,j} &= \sqrt{(d_{x_1})\mathbb{1}_{y_{x_1}=j}} - \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\ &= \sqrt{d_{x_1}} - \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\ &= \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}} - \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\ &= \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq j} \\ &= \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}}.\end{aligned} \tag{32}$$

When $j \neq y_{x_1}$, $(D^{1/2}Y - \bar{A}C)_{x_1,j} = \sqrt{(d_{x_1})\mathbb{1}_{y_{x_1}=j}} - \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j}$

$$\begin{aligned}&= 0 - \sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\ &= -\sum_{x_1^+}\frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j}.\end{aligned} \tag{33}$$

We define $\beta_{x_1} = \sum_{x_1^+} \mathcal{A}(x_1, x_1^+) \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}$, and we have

$$\begin{aligned}
\|(D^{1/2}Y - \bar{A}C)_{x_1}\|^2 &= \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}\right)^2 + \sum_{j \neq y_{x_1}} \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} = j}\right)^2 \\
&\leq \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}\right)^2 + \left(\sum_{j \neq y_{x_1}} \sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} = j}\right)^2 \\
&\leq \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}\right)^2 + \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \sum_{j \neq y_{x_1}} \mathbb{1}_{y_{x_1^+} = j}\right)^2 \\
&= \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}\right)^2 + \left(\sum_{x_1^+} \frac{\mathcal{A}(x_1, x_1^+)}{\sqrt{d_{x_1}}} \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}\right)^2 \\
&= \frac{2\beta_{x_1}^2}{d_{x_1}}.
\end{aligned} \tag{34}$$

With that, we obtain $\|D^{1/2}Y - \bar{A}C\|^2 = \sum_{x_1} \frac{2\beta_{x_1}^2}{d_{x_1}}$. Let $\beta = \sum_{x_1, x_1^+} \mathcal{A}(x_1, x_1^+) \mathbb{1}_{y_{x_1^+} \neq y_{x_1}}$. Then we have

$$\begin{aligned}
\|D^{1/2}Y - \bar{A}C\|^2 &= \sum_{x_1} \frac{2\beta_{x_1}^2}{d_{x_1}} \\
&\leq \sum_{x_1} 2\beta_{x_1} \qquad \left(d_{x_1} = \sum_{x_1^+} \mathcal{A}(x_1, x_1^+) \geq \beta_{x_1}\right) \\
&= 2\beta.
\end{aligned}$$

Then we obtain

$$\begin{aligned}
\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 &= \|D^{1/2}Y - \bar{A}C + \bar{A}C - UW_f^\top\|^2 \\
&\leq 2\|\bar{A}C - UW_f^\top\|^2 + 4\beta \\
&= 2\|(\bar{A} - UU^T + UU^T)C - UW_f^\top\|^2 + 4\beta \\
&= 2\|(\bar{A} - UU^T)C + U(U^TC - W_f^\top)\|^2 + 4\beta \\
&\leq 4\big(\|(\bar{A} - UU^T)C\|^2 + \|U(U^TC - W_f^\top)\|^2\big) + 4\beta \\
&\leq 4\big(\|\bar{A} - UU^T\|^2\|C\|^2 + \|U\|^2\|U^TC - W_f^\top\|^2\big) + 4\beta \\
&\leq 4\big(\|\bar{A} - UU^T\|^2 + \|U^TC - W_f^\top\|^2\big) + 4\beta \\
&= 4\mathcal{L}_{SCL}(f) + \text{const} + \|U^TC - W_f^\top\|^2 + 4\beta \\
&\leq 16L \cdot \mathcal{L}_{\text{U-MPT}}(h) + 8\mathbb{E}_y[\mathbb{E}_{x|y} f(x) - b_y]^2 + 4\beta + 16L\varepsilon + \text{const} \\
&\leq 16L \cdot \mathcal{L}_{\text{U-MPT}}(h) + 4\beta + 16L\varepsilon + \text{const}.
\end{aligned} \tag{35}$$

Here we leverage Lemma B.8 in HaoChen et al. (2021) and the fact that W_f is a mean classifier. In the next step, we analyze the prediction error. We denote $y(\bar{x})$ as the ground-truth label of original data $\bar{x}$. We first define an ensembled linear predictor c'_f. For an original sample, the predictor ensembles the results of all different views and choose the label predicted most. With the definition, $y(\bar{x}) \neq c'_f(\bar{x})$ only happens when more than half of the views predict wrong labels. So

$$\begin{aligned}
&\Pr(y(\bar{x}) \neq c'_f(\bar{x})) \\
&\leq 2\Pr(y(\bar{x}) \neq c_f(x)) \\
&\leq 4\mathbb{E}_{\bar{x}\sim\mathcal{P}_d(x),x\sim\mathcal{M}_1(x|\bar{x})}\|y(\bar{x}) - W_f f(x)\|^2 \\
&\leq 8(\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 + \mathbb{E}_{\bar{x}\sim\mathcal{P}_d(x),x\sim\mathcal{M}_1(x|\bar{x})}\|y - y(\bar{x})\|^2) \\
&\leq 8(\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 + 2\alpha) \\
&\leq 32\mathcal{L}_{SCL}(f) + 64\alpha + 16\alpha + \text{const} \\
&\leq 128L \cdot \mathcal{L}_{\text{U-MPT}}(h) + 32\beta + 16\alpha + 128L\varepsilon + \text{const}.
\end{aligned} \tag{36}$$

Also note that, as we assume $E_{\bar{x}\sim\mathcal{P}_d}(\mathcal{A}(x_1|\bar{x})\mathbb{1}_{y_{x_1}\neq y(\bar{x})}) \leq \alpha$, we have

$$\begin{aligned}
\beta &= \sum_{x_1,x_1^+} E_{\bar{x}}(\mathcal{A}(x_1|\bar{x})\mathcal{A}(x_{x_1^+}|\bar{x})\mathbb{1}_{y_{x_1^+}\neq y_{x_1}}) \\
&\leq \sum_{x_1,x_1^+} E_{\bar{x}}(\mathcal{A}(x_1|\bar{x})\mathcal{A}(x_{x_1^+}|\bar{x})(\mathbb{1}_{y_i\neq y(\bar{x})} + \mathbb{1}_{y_{x_1^+}\neq y(\bar{x})})) \\
&= 2E_{\bar{x}\sim\mathcal{P}_d}(\mathcal{A}(x_1|\bar{x})\mathbb{1}_{y_{x_1}\neq y(\bar{x})}) \\
&\leq 2\alpha.
\end{aligned} \tag{37}$$

This implies

$$\Pr(y(\bar{x}) \neq c'_f(\bar{x})) \leq 128L \cdot \mathcal{L}_{\text{U-MPT}}(h) + 80\alpha + 128L\varepsilon + \text{const},$$

which gives the values of the constants mentioned in Theorem 9.

■

A.6 Proof of Theorem 10

Proof With Equation (35), we have

$$\mathcal{L}_{\text{U-MPT}}(h) \geq \frac{1}{4L}\|A - UU^\top\|^2 - \varepsilon + \text{const}. \tag{38}$$

We set $\mathcal{L}_{mf}(U) = \|(\bar{A} - UU^T)\|^2$. When $U^\star$ is the minimizer of $\mathcal{L}_{mf}(U)$, according to the analysis in Eckart and Young (1936), we obtain

$$\|(\bar{A} - U^\star(U^\star)^T)\|^2 = \sum_{i=k+1}^{N_1} \lambda_i^2, \tag{39}$$

where $\lambda_{k+1} \cdots \lambda_{N_1}$ are the remaining $N_1 - k$ eigenvalues of matrix $\bar{A}$. We denote $h^\star$ as the minimizer of $\mathcal{L}_{\text{U-MPT}}$ and $f^\star$ as the corresponding encoder. Then $U_{f^\star}$ is composed of the features encoded by $f^\star$, *i.e.,* $(U_{f^\star})_{x_1} = \sqrt{d_{x_1}} f^\star(x_1)$. So for all $h \in \mathcal{H}$, we have

$$\begin{aligned} \mathcal{L}_{\text{U-MPT}}(h) \geq \mathcal{L}_{\text{U-MPT}}(h^\star) &\geq \frac{1}{4L}\|A - U_{f^\star}(U_{f^\star})^\top\|^2 - \varepsilon + \text{const} \\ &\geq \frac{1}{4L}\|A - U^\star(U^\star)^\top\|^2 - \varepsilon + \text{const} \\ &= \frac{1}{4L}\sum_{i=k+1}^{N_1} \lambda_i^2 - \varepsilon + \text{const}. \end{aligned} \tag{40}$$

■

Appendix B. Analysis of the Toy Model

In this section, we provide the theoretical proofs of the toy model.

B.1 Problem Setting and Probabilistic Model

Consider two classes, C_1 and C_2, and K patch positions arranged on a two-dimensional grid. For each class C_i and position k, define a local candidate set

$$P_{i,k} = E_{i,k} \cup O_k,$$

where $E_{i,k}$ contains N class-specific patches and O_k contains M patches shared by the two classes. At the same position, the class-specific sets are disjoint, i.e., $E_{1,k} \cap E_{2,k} = \emptyset$, and the two classes overlap only through O_k. The identities of these candidate patches may vary with k. If an image belongs to class C_i, then its k-th patch is independently and uniformly sampled from $P_{i,k}$.

To create a masked view from an image, ρK positions are randomly selected as the masked portion, while the remaining $(1-\rho)K$ positions form the unmasked view, where ρ is the mask ratio. Let I_2 denote the masked-position set and I_1 denote the unmasked-position set. A view is termed an overlapped view if all its patches are drawn from the shared local sets O_k at the corresponding positions; otherwise, it is considered a non-overlapped view.

The following quantities come in handy:

$$\begin{aligned} H &= \prod_{k \in I_2} |P_{i,k}| = (N+M)^{\rho K}, \\ W &= \prod_{k \in I_2} |O_k| = M^{\rho K}, \\ R &= \prod_{k \in I_1} |P_{i,k}| = (N+M)^{(1-\rho)K}, \\ G &= \prod_{k \in I_1} |O_k| = M^{(1-\rho)K}. \end{aligned}$$

where H, W, R, G denote the number of possible masked views, overlapped masked views, possible unmasked views, and overlapped unmasked views, respectively.

Given the uniform sampling, the marginal probability of a view with p patches being sampled from a specific class is $\frac{1}{(N+M)^p}$. Considering both classes, we can derive the marginal probability of each view as follows:

- Non-overlapped masked views: $\frac{1}{2H}$
- Overlapped masked views: $\frac{1}{H}$
- Non-overlapped unmasked views: $\frac{1}{2R}$
- Overlapped unmasked views: $\frac{1}{R}$

Note that overlapped views have twice the probability of being selected as non-overlapped views, since they can be chosen by both classes.

We use x_1, x_1^+ to denote unmasked views, and x_2 masked views. The quantity w_{x_1,x_2} represents the probability that x_1 and x_2 come from the same image. Specifically, $w_{x_1,x_2} = \Pr(x_2)\Pr(x_1|x_2)$. Clearly, $w_{x_1,x_2} = 0$ if both x_1 and x_2 are non-overlapped and from different classes. Otherwise, w_{x_1,x_2} is equal to $\frac{1}{2HR}$ when x_1 or x_2 is non-overlapped, and $\frac{1}{HR}$ when both are overlapped.

The nodes in an augmentation graph correspond to the set of all unmasked views. The normalized edge weight between two unmasked views, x_1 and x_1^+, is given by

$$\mathcal{A}^*(x_1, x_1^+) = \frac{1}{\sqrt{w_{x_1} w_{x_1^+}}} \sum_{x_2} \frac{w_{x_1,x_2} w_{x_1^+,x_2}}{w_{x_2}}$$

where w_{x_1} represents the marginal probability of x_1. Here we use a normalized version of the augmentation graph for the sake of simplicity.

$\mathcal{A}^*(x_1, x_1^+)$ can take one of the following values:

- **One of x_1 and x_1^+ is non-overlapped and the other is overlapped**: $\mathcal{A}^*(x_1, x_1^+) = \frac{1}{\sqrt{2}R} := p_1$.
- **x_1 and x_1^+ are in the same class and both non-overlapped**: $\mathcal{A}^*(x_1, x_1^+) = \frac{2H-W}{2HR} := p_2$.
- **x_1 and x_1^+ are in different classes and both non-overlapped**: $\mathcal{A}^*(x_1, x_1^+) = \frac{W}{2HR} := p_3$.
- **Both x_1 and x_1^+ are overlapped**: $\mathcal{A}^*(x_1, x_1^+) = \frac{1}{R} := p_4$.

B.2 Adjacency Matrix and Graph Connectivity

The adjacency matrix of the augmentation graph, A, is symmetric and structured into 9 blocks, with the widths of blocks labeled above:

$$A = \left[\begin{array}{c|c|c} \overbrace{A_{11}}^{R-G} & \overbrace{A_{12}}^{R-G} & \overbrace{A_{13}}^{G} \\ \hline A_{21} & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \end{array}\right].$$

Each block A_{ij} $(i = 1, 2, 3; j = 1, 2, 3)$ is a constant matrix, and the entries are given below:

$$C = (c_{ij})_{3\times 3} = \begin{bmatrix} p_2 & p_3 & p_1 \\ p_3 & p_2 & p_1 \\ p_1 & p_1 & p_4 \end{bmatrix}.$$

where c_{ij} is the value of the entries in the block A_{ij}.

A block A_{ij} can be expressed as $A_{ij} = c_{ij}\mathbb{1}_{n_i}\mathbb{1}_{n_j}^\top$, where $(n_1, n_2, n_3) = (R-G, R-G, G)$ and $\mathbb{1}_n$ denotes a column vector of ones with dimension n.

Therefore,

$$A_1 = \begin{bmatrix} c_{11}\mathbb{1}_{n_1}\mathbb{1}_{n_1}^\top & c_{12}\mathbb{1}_{n_1}\mathbb{1}_{n_2}^\top & c_{13}\mathbb{1}_{n_1}\mathbb{1}_{n_3}^\top \\ c_{21}\mathbb{1}_{n_2}\mathbb{1}_{n_1}^\top & c_{22}\mathbb{1}_{n_2}\mathbb{1}_{n_2}^\top & c_{23}\mathbb{1}_{n_2}\mathbb{1}_{n_3}^\top \\ c_{31}\mathbb{1}_{n_3}\mathbb{1}_{n_1}^\top & c_{32}\mathbb{1}_{n_3}\mathbb{1}_{n_2}^\top & c_{33}\mathbb{1}_{n_3}\mathbb{1}_{n_3}^\top \end{bmatrix} = \left[\begin{bmatrix} c_{11}\mathbb{1}_{n_1} \\ c_{12}\mathbb{1}_{n_2} \\ c_{13}\mathbb{1}_{n_3} \end{bmatrix}\mathbb{1}_{n_1}^\top \quad \begin{bmatrix} c_{21}\mathbb{1}_{n_1} \\ c_{22}\mathbb{1}_{n_2} \\ c_{23}\mathbb{1}_{n_3} \end{bmatrix}\mathbb{1}_{n_2}^\top \quad \begin{bmatrix} c_{31}\mathbb{1}_{n_1} \\ c_{32}\mathbb{1}_{n_2} \\ c_{33}\mathbb{1}_{n_3} \end{bmatrix}\mathbb{1}_{n_3}^\top\right].$$

It is evident that any vector in the nullspace of $\mathbb{1}_{n_i}^\top$ can be expanded to a vector in the nullspace of A, by filling the remaining blocks with zeros. Therefore, A has an eigenvalue of 0 with multiplicity at least $2R - G - 3$.

In order to find the other three eigenvalues, we further consider the matrix Q:

$$Q = (q_{ij})_{3\times 3} = \begin{bmatrix} (R-G)p_2 & (R-G)p_3 & Gp_1 \\ (R-G)p_3 & (R-G)p_2 & Gp_1 \\ (R-G)p_1 & (R-G)p_1 & Gp_4 \end{bmatrix}.$$

where $q_{ij} = n_j c_{ij}$. Let λ be an eigenvalue of Q, with corresponding eigenvector $\boldsymbol{v}$. From the equation $Q\boldsymbol{v} = \lambda\boldsymbol{v}$, we have

$$\sum_{j=1}^{3} c_{ij} n_j v_j = \lambda v_i,$$

where v_i is the i-th element of $\boldsymbol{v}$. Next, we define $\boldsymbol{u} = \boldsymbol{v} \otimes \begin{bmatrix} \mathbb{1}_{n_1} \\ \mathbb{1}_{n_2} \\ \mathbb{1}_{n_3} \end{bmatrix} = \begin{bmatrix} v_1\mathbb{1}_{n_1} \\ v_2\mathbb{1}_{n_2} \\ v_3\mathbb{1}_{n_3} \end{bmatrix}$, where $\otimes$ denotes the Kronecker product. Let $\boldsymbol{w} = A\boldsymbol{u}$. Then, the i-th block of $\boldsymbol{w}$, $\boldsymbol{w}_i$, satisfies:

$$\boldsymbol{w}_i = \sum_{j=1}^{3} A_{ij}\boldsymbol{u}_j = \sum_{j=1}^{3} c_{ij}\mathbb{1}_{n_i}\mathbb{1}_{n_j}^\top v_j \mathbb{1}_{n_j} = \sum_{j=1}^{3}(c_{ij}n_j v_j)\mathbb{1}_{n_i} = \sum_{j=1}^{3}\lambda v_i \mathbb{1}_{n_i} = \sum_{j=1}^{3}\lambda \boldsymbol{u}_i,$$

Thus, $\boldsymbol{u}$ is an eigenvector of A with corresponding eigenvalue λ. From the structure of $\boldsymbol{u}$, we observe that $\boldsymbol{u}$ is orthogonal to all eigenvectors corresponding to the eigenvalue 0 discussed above. Finally, we can find the remaining three eigenvalues as below:

$$\begin{cases} \lambda_1 & = 1, \\ \lambda_2 & = \frac{(H-W)(R-G)}{HR} = \left[1 - \left(\frac{M}{N+M}\right)^{\rho K}\right]\left[1 - \left(\frac{M}{N+M}\right)^{(1-\rho)K}\right], \\ \lambda_3 & = 0. \end{cases}$$

We observe that λ_2 lies within the range $(0, 1)$ and remains unchanged if ρ is replaced by $1 - \rho$, reflecting a symmetry between the masked and unmasked views. The value of λ_2 is maximized at $\rho = 0.5$, and diminishes as ρ moves away from this midpoint. According to spectral graph theory Chung (1997), a smaller λ_2 implies a more connected augmentation graph, supporting our analysis in Section 6.2 that connectivity improves when the mask ratio approaches 1.

B.3 The Label Error

Denote the unmasked-to-unmasked label error as $\beta = \sum_{x_1, x_1^+} \mathcal{A}(x_1, x_1^+)\mathbb{1}[y(x_1) \neq y(x_1^+)]$. To evaluate β, we consider the following four scenarios:

- **One of x_1 and x_1^+ is non-overlapped and the other is overlapped**: $\mathcal{A}(x_1, x_1^+) = \frac{1}{2R^2} = q_1$. In this case, $\Pr[y(x_1) \neq y(x_1^+)] = \eta$.
- **x_1 and x_1^+ are in the same class and both non-overlapped**: $\mathcal{A}(x_1, x_1^+) = \frac{2H-W}{4HR^2} = q_2$. In this case, $\Pr[y(x_1) \neq y(x_1^+)] = 0$.
- **x_1 and x_1^+ are in different classes and both non-overlapped**: $\mathcal{A}(x_1, x_1^+) = \frac{W}{4HR^2} = q_3$. In this case, $\Pr[y(x_1) \neq y(x_1^+)] = 1$.
- **Both x_1 and x_1^+ are overlapped**: $\mathcal{A}(x_1, x_1^+) = \frac{1}{R^2} = q_4$. In this case, $\Pr[y(x_1) \neq y(x_1^+)] = \eta$.

Here, η should be a value greater than 0 and considerably smaller than 0.5. The reason is that in practice, even if a patch possibly belongs to both two classes, its membership to a class should still be on a spectrum, *i.e.,* its belonging is not completely indistinguishable; considering that an unmasked view comprises many patches, the indistinguishability, *i.e.,* label error should be significantly smaller than the completely stochastic scenario. So, we assume $0 < \eta < 0.3$ here.

Then we have

$$\begin{aligned} \beta &= 4(R-G)G \cdot q_1 \cdot \eta + 2(R-G)^2 \cdot q_2 \cdot 0 + 2(R-G)^2 \cdot q_3 \cdot 1 + G^2 \cdot q_4 \cdot \eta \\ &= \frac{8H(R-G)G\eta + 2(R-G)^2 W + 4HG^2\eta}{4HR^2}. \end{aligned}$$

Denote the unmasked-to-natural label error as $\alpha = \mathbb{E}_{\bar{x}, x_1}\mathbb{1}[y(x_1) \neq y(\bar{x})]$. We have

$$
\begin{aligned}
\alpha &= \mathbb{E}_{\bar{x}}\Pr[y(x_1) \neq y(\bar{x})] \\
&= \Pr(x_1 \text{ is overlapped})[\Pr(x_1 \in C_2 | x_1 \text{ is overlapped})\Pr(\bar{x} \in C_1 | x_1 \text{ is overlapped}) \\
&\qquad\qquad + \Pr(x_1 \in C_1 | x_1 \text{ is overlapped})\Pr(\bar{x} \in C_2 | x_1 \text{ is overlapped})] \\
&= \frac{G}{R}\left[\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\right] \\
&= \frac{G}{2R} \\
&= \left(\frac{M}{N+M}\right)^{(1-m)K}.
\end{aligned}
$$

B.4 Proof of Theorem 11

Proof As shown in Equation (36), we consider $S = 32\lambda_2^2 + 32\beta + 16\alpha$ as a component of the upper bound for the downstream classification error that relates to ρ. Let $w = \frac{W}{H} = r^{\rho K}$, $g = \frac{G}{R} = r^{(1-\rho)K}$ and $V = wg = r^K$. Note that V is independent of the mask ratio ρ. As stated in Theorem 11, V satisfies $0 < V < \frac{1}{12}$. Then we have

$$
\begin{aligned}
S &= 32\lambda_2^2 + 32\beta + 8\alpha \\
&= 32\left[\frac{(H-W)(R-G)}{HR}\right]^2 + 32\frac{8H(R-G)G\eta + 2(R-G)^2W + 4HG^2\eta}{4HR^2} + 8\frac{G}{2R} \\
&= 32g^2w^2 - 48g^2w + 32(1-\eta)g^2 - 64gw^2 + 96gw + [64(\eta-1)+4]g + 32w^2 - 48w + 32.
\end{aligned}
$$

Let $S^*(\rho, r, K) = \frac{S-32}{4}$. We now regard g and w as variables that comply with the constraints $V < w < 1$ and $wg = V$. By plugging in $wg = V$ we get

$$
\begin{aligned}
S^* &= 8V^2 - 12Vg + 8(1-\eta)g^2 - 16Vw + 24V + [16(\eta-1)+1]g + 8w^2 - 12w \\
&= 8(1-\eta)g^2 + [16(\eta-1) + 1 - 12V]g + 8w^2 - (16V+12)w + 8V^2 + 24V.
\end{aligned}
$$

Let $x = 2\sqrt{2}\cdot w$, $y = 2\sqrt{2(1-\eta)}\cdot w$, $a = \frac{\sqrt{2}(4V+3)}{2}$ and $b = \frac{12V-16\eta+15}{4\sqrt{2(1-\eta)}}$. Then $S^* = (x-a)^2 + (y-b)^2 + \Phi(V,\eta)$, where $\Phi(V,\eta)$ is a function of η and V which is independent of w and g, and consequently irrelevant to ρ. We also note that $xy = 8\sqrt{1-\eta}\cdot V := c$. By plugging this in, we have

$$
S^* = (x-a)^2 + \left(\frac{c}{x} - b\right)^2 + \Phi(V,\eta).
$$

Take the derivative of S^* with respect to x:

$$
\frac{\partial S^*}{\partial x} = 2(x-a) - \frac{2c}{x^2}\left(\frac{c}{x} - b\right)
$$

We consider the sign of the derivative at two points.
When $x = \frac{c}{b}$,

$$
\begin{aligned}
\left.\frac{\partial S^*}{\partial x}\right|_{x=\frac{c}{b}} &= 2\left(\frac{c}{b}-a\right) \\
&= \frac{2}{b}(c-ab) \\
&= \frac{2}{b}\left[8\sqrt{1-\eta}\cdot V - \frac{(4V+3)\sqrt{2}}{2}\frac{12V-16\eta+15}{4\sqrt{2(1-\eta)}}\right] \\
&= \frac{2}{b}\cdot\frac{-48V^2-32V+48\eta-45}{8\sqrt{1-\eta}} \\
&< \frac{2}{b}\cdot\frac{48\cdot 0.3-45}{8\sqrt{1-\eta}} && (V>0,\ \eta<0.3) \\
&< 0.
\end{aligned}
$$

When $x=\sqrt{c}$,

$$
\begin{aligned}
\left.\frac{\partial S^*}{\partial x}\right|_{x=\sqrt{c}} &= 2\sqrt{c}(b-a) \\
&= 2\sqrt{c}\cdot\left[\frac{12V-16\eta+15}{4\sqrt{2(1-\eta)}} - \frac{(4V+3)\sqrt{2}}{2}\right] \\
&= \frac{\sqrt{2c}}{4\sqrt{1-\eta}}\cdot\left[-4V\left(4\sqrt{1-\eta}-3\right)-16\eta-12\sqrt{1-\eta}+15\right] \\
&> \frac{\sqrt{2c}}{4\sqrt{1-\eta}}\cdot\left[16-16\eta-\frac{40}{3}\sqrt{1-\eta}\right] && \left(V<\frac{1}{12}\right) \\
&= \sqrt{2c}\cdot\left[16\sqrt{1-\eta}-\frac{40}{3}\right] \\
&> \sqrt{2c}\cdot\left[16\sqrt{0.7}-\frac{40}{3}\right] && (\eta<0.3) \\
&> 0. && (41)
\end{aligned}
$$

From the signs of the derivative at certain points we know that S^* reaches a local minimum at $x=x^*$, where $\frac{c}{b}<x^*<\sqrt{c}$ (if there are multiple local minima, we take the one that leads to the smallest S^*). Also note that $\frac{c}{b}<\sqrt{c}$ holds, since this is equivalent to $c<b^2$ and we have

$$
\begin{aligned}
c &= 8\sqrt{1-\eta}\cdot V \\
&< 0.67 && \left(V < \frac{1}{12},\ \eta > 0\right) \\
&< 1.8 \\
&< \frac{12V - 16\eta + 15}{4\sqrt{2(1-\eta)}} && (V > 0,\ 0 < \eta < 0.3) \\
&= b \\
&< b^2.
\end{aligned}
$$

$x < \sqrt{c}$ is equivalent to $w = \frac{x}{2\sqrt{2}} < \sqrt{\sqrt{(1-\eta)}V} < \sqrt{V}$; together with $w = r^{\rho K} = V^{\rho}$ and $V = r^K < 1$ we have $\rho > 0.5$. $x > \frac{c}{b}$ is equivalent to $w > \frac{c}{2\sqrt{2}\cdot b}$, which means

$$
\begin{aligned}
w &> \frac{8\sqrt{1-\eta}}{2\sqrt{2}} \cdot \frac{4\sqrt{2(1-\eta)}}{12V - 16\eta + 15} \cdot V \\
&= \frac{16(1-\eta)}{12V + 15 - 16\eta} \cdot V \\
&> \frac{16(1-\eta)}{16 - 16\eta} \cdot V && \left(V < \frac{1}{12}\right) \\
&= V.
\end{aligned}
$$

This in turn implies $\rho < 1$. Therefore, the local minimum x^* is achieved when $0.5 < \rho < 1$.

It remains to show that this local minimum is a global minimum when $w > 0$, *i.e.*, $x > 0$. Consider the second derivative of S^* with respect to x:

$$
\frac{\partial^2 S^*}{\partial x^2} = \frac{2c^2}{x^4} + \frac{4c\left(\frac{c}{x} - b\right)}{x^3} + 2.
$$

Clearly, $\frac{\partial^2 S^*}{\partial x^2} > 0$ when $x < \frac{c}{b}$. This implies that S^* does not have a zero point when $x < \frac{c}{b}$, meaning that x^* is a global minimum for $0 < x < \sqrt{c}$. For $x > \sqrt{c}$, we consider the symmetry of the hyperbola $xy = C$. Let $x_1 < \sqrt{c}$, then $\frac{c}{x_1} > \sqrt{c}$. We have

$$
\begin{aligned}
S^*|_{x=x_1} - S^*|_{x=\frac{c}{x_1}} &= (x_1 - a)^2 + \left(\frac{c}{x_1} - b\right)^2 - \left(\frac{c}{x_1} - a\right)^2 - (x_1 - b)^2 \\
&= \left(x_1 + \frac{c}{x_1} - 2a\right)\left(x_1 - \frac{c}{x_1}\right) - \left(x_1 + \frac{c}{x_1} - 2b\right)\left(x_1 - \frac{c}{x_1}\right) \\
&= 2\left(x_1 - \frac{c}{x_1}\right)(b - a) \\
&< 0. && \left(x_1 < \sqrt{c} < \frac{c}{x_1};\ b - a > 0, \text{ from inequality 41}\right) \quad (42)
\end{aligned}
$$

Now, suppose there exists a local minimum $\hat{x} > \sqrt{c}$. From inequality (42), we know that $S^*|_{x=\hat{x}} > S^*|_{x=\frac{c}{\hat{x}}}$. Since x^* is a global minimum for $0 < x < \sqrt{c}$ and $0 < \frac{c}{\hat{x}} < \sqrt{c}$, we

have $S^*|_{x=\frac{c}{\hat{x}}} > S^*|_{x=x^*}$. Thus, any local minimum $\hat{x} > \sqrt{c}$ results in greater S^* than x^* does.

In conclusion, x^* is the global minimum for $x > 0$. This completes the proof of Theorem 11. ■

Appendix C. Visualization of the Augmentation Graph

To provide a better visualization of the augmentation graph, we randomly selected 6 image pairs from the ImageNet data set: 3 intra-class pairs and 3 inter-class pairs. For each image, 500 random masks were applied with a mask ratio of 75%. A Vision Transformer, pretrained using Masked Autoencoder (MAE) on ImageNet, was then utilized to estimate the similarity between each pair of unmasked views. This estimation was based on the cosine similarity of the extracted features. From these, the top three unmasked view pairs, as determined by their similarity scores, were selected for visualization. The results are presented in Figure 8.

Clearly, unmasked views within intra-class image pairs exhibit significantly higher similarity compared to those in inter-class pairs, which also leads to larger weights for the edges in the augmentation graph. The similarity is illustrated by the presence of common iconic patches, such as the wheels in Figure 8(a), the fur in Figure 8(b) and the piles of oranges in Figure 8(c). Conversely, the unmasked regions of inter-class image pairs appear more heterogeneous and lack a consistent pattern, underscoring the fundamental visual differences between distinct classes.

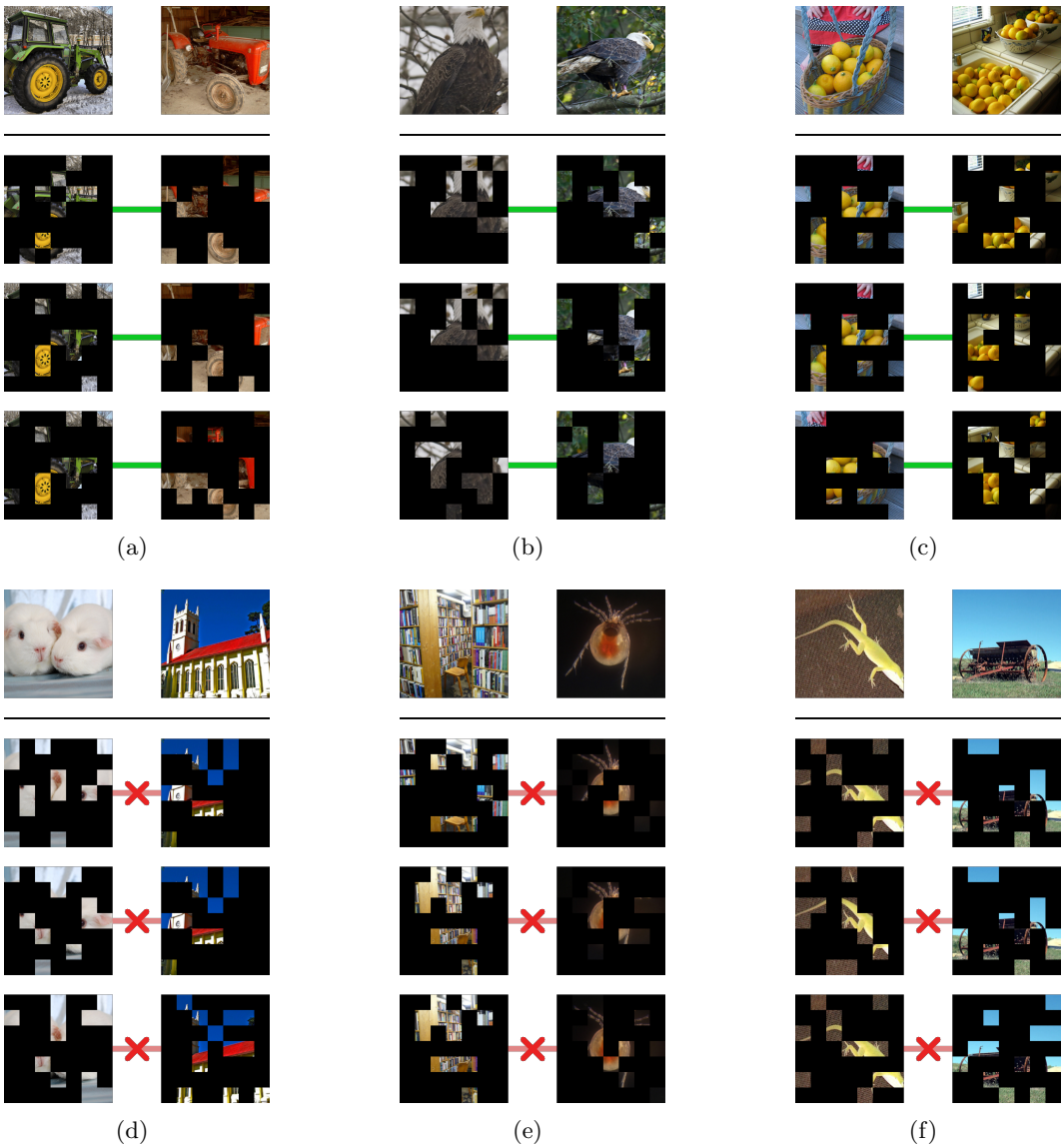


Figure 8: Illustration of edges in an augmentation graph between intra-class images and inter-class images. The first row shows that intra-class image pairs tend to have fairly overlapped unmasked views, while the second row demonstrates that even the most similar unmasked views of inter-class image pairs are non-overlapped.

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Jingyi Cui, Weiran Huang, Yifei Wang, and Yisen Wang. Rethinking weak supervision in helping contrastive learning. In *ICML*, 2023.

Jingyi Cui, Hongwei Wen, and Yisen Wang. An augmentation-aware theory for self-supervised contrastive learning. In *ICML*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

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. In *ICLR*, 2021.

Tianqi Du, Yifei Wang, and Yisen Wang. On the role of discrete tokenization in visual representation learning. In *ICLR*, 2024.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: a new approach to self-supervised learning. In *NeurIPS*, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *CVPR*, 2021.

Zhicheng Huang, Xiaojie Jin, Cheng Lu, Qibin Hou, Mingg-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2506–2517, 2024.

Ziyu Jiang, Yinpeng Chen, Mengchen Liu, Dongdong Chen, Xiyang Dai, Lu Yuan, Zicheng Liu, and Zhangyang Wang. Layer grafted pre-training: Bridging contrastive learning and masked image modeling for label-efficient representations. In *ICLR*, 2023.

Li Jing, Jure Zbontar, and Yann LeCun. Implicit rank-minimizing autoencoder. In *NeurIPS*, 2020.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2022.

Zaid Khan and Yun Fu. Exploiting bert for multimodal target sentiment classification through input space translation. In *ACM MM*, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

Lingjing Kong, Martin Q. Ma, Guan-Hong Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *CVPR*, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Yuchen Li, Alexandre Kirchmeyer, Aashay Mehta, Yilong Qin, Boris Dadachev, Kishore Papineni, Sanjiv Kumar, and Andrej Risteski. Promises and pitfalls of generative masked language modeling: Theoretical framework and practical guidelines. In *ICML*, 2024.

Bingbin Liu, Daniel J Hsu, Pradeep Ravikumar, and Andrej Risteski. Masked prediction: A parameter identifiability view. In *NeurIPS*, 2022.

Zhengqi Liu, Jie Gui, and Hao Luo. Good helper is around you: Attention-driven masked image modeling. In *AAAI*, 2023.

Yu Meng, Jitin Krishnan, Sinong Wang, Qifan Wang, Yuning Mao, Han Fang, Marjan Ghazvininejad, Jiawei Han, and Luke Zettlemoyer. Representation deficiency in masked language modeling. In *ICLR*, 2024.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. In *ICLR*, 2025.

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 *ICML*, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *EUSIPCO*, 2007.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, 2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *CVPR*, 2023a.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Yifei Wang, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Residual relaxation for multi-view representation learning. In *NeurIPS*, 2021.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *ICLR*, 2022.

Yifei Wang, Qi Zhang, Tianqi Du, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. A message passing perspective on learning dynamics of contrastive learning. In *ICLR*, 2023b.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. Mrm: Masked relation modeling for medical image pre-training with genetics. In *ICCV*, 2023.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *ICLR*, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *NeurIPS*, 2022a.

Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In *ICML*, 2023.

Qi Zhang, Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. Look ahead or look around? a theoretical comparison between autoregressive and masked pretraining. In *ICML*, 2024.

Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. Beyond interpretability: The gains of feature monosemanticity on model robustness. In *ICLR*, 2025a.

Qi Zhang, Yifei Wang, and Yisen Wang. An augmentation overlap theory of contrastive learning. *Journal of Machine Learning Research*, 26(228):1–42, 2025b.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J Passonneau. Contrastive data and learning for natural language processing. In *NAACL*, 2022b.

Yi-Ge Zhang, Jingyi Cui, Qiran Li, and Yisen Wang. Difficult examples hurt unsupervised contrastive learning: A theoretical perspective. In *ICLR*, 2026.

Zhijian Zhuo, Yifei Wang, Jinwen Ma, and Yisen Wang. Towards a unified theoretical understanding of non-contrastive learning via rank differential mechanism. In *ICLR*, 2023.