arXiv:submit/4546223 [cs.LG] 15 Oct 2022

How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders

Qi Zhang[1*] **Yifei Wang** [2*] **Yisen Wang**[1,3†]
[1] Key Lab. of Machine Perception (MoE),
School of Intelligence Science and Technology, Peking University
[2] School of Mathematical Sciences, Peking University
[3] Institute for Artificial Intelligence, Peking University

Abstract

Masked Autoencoders (MAE) based on a reconstruction task have risen to be a promising paradigm for self-supervised learning (SSL) and achieve state-of-the-art performance across different benchmark datasets. However, despite its impressive empirical success, there is still limited theoretical understanding of it. In this paper, we propose a theoretical understanding of how masking matters for MAE to learn meaningful features. We establish a close connection between MAE and contrastive learning, which shows that MAE implicit aligns the mask-induced positive pairs. Built upon this connection, we develop the first downstream guarantees for MAE methods, and analyze the effect of mask ratio. Besides, as a result of the implicit alignment, we also point out the dimensional collapse issue of MAE, and propose a Uniformity-enhanced MAE (U-MAE) loss that can effectively address this issue and bring significant improvements on real-world datasets, including CIFAR-10, ImageNet-100, and ImageNet-1K. Code is available at https://github.com/zhangq327/U-MAE.

1 Introduction

Recently, self-supervised learning (SSL) has been proposed as a promising paradigm for learning data representations without access to labels. Beside the popular contrastive learning methods [4, 14, 12, 27], recently there has been a renaissance of reconstruction-based autoencoders for SSL, for example, MAE [15], BEiT [1], iBOT [32], and SimMIM [29], which demonstrate state-of-the-art performance on various downstream tasks. Taking MAE as an example, it learns to reconstruct the masked patches from the unmasked context with an encoder-decoder architecture, from which we can see that there are two main components in MAE: masking and autoencoders. While for autoencoders, its canonical one, even equipped with expressive neural networks, is still less competitive than modern SSL methods [4], which indicates that autoencoders may not be the key factor for the success of MAE. Then for masking, MAE needs a very large mask ratio, *e.g.,* 75% patches of the input image are masked out. Such a large mask ratio will lead to the loss of most image contents that are hardly recoverable from the rest. Further considering that MAE deliberately chooses a weak decoder with a shadow architecture, we believe that reconstruction might not be the ultimate goal of MAE in the learning process. Then, some natural questions are raised here:

What is the role of masking in MAE? How does it affect the downstream performance?

To answer these questions, firstly, we build a close connection between MAE and contrastive learning. In particular, we find that masking could produce *implicit* positive pairs while MAE's reconstruction

[*]Equal Contribution.
[†]Corresponding author: Yisen Wang (yisen.wang@pku.edu.cn).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

loss is lower bounded by an *alignment* loss on these positive pairs. Further for feature uniformity, we notice that although MAE will not fully collapse, it still suffers from *dimensional collapse* where features lie in a low-dimensional space and become highly alike. Inspired by the uniformity loss in contrastive learning, we propose a Uniformity-enhanced MAE (**U-MAE**) to promote the feature diversity. Empirically, the proposed U-MAE attains significant improvements over MAE on the linear probing task across different benchmark datasets (CIFAR-10, ImageNet-100, and ImageNet-1K) and different ViT backbones, and it effectively eliminates the feature collapse issue of MAE. Secondly, the connection between MAE and contrastive learning enables us to establish the *first theoretical guarantee* for downstream classification among MAE methods. Our guarantees suggest that a large mask ratio is indeed necessary for bridging semantically similar samples together.

We summarize our contributions as follows:

- We propose a new theoretical understanding of MAE by establishing a formal connection between MAE and contrastive learning. In particular, we show that a small reconstruction loss implies better alignment of mask-induced positive pairs.
- Built upon this connection, we establish the first theoretical guarantee on the downstream performance among MAE methods, which helps to understand the role of high mask ratio.
- We point out the dimensional collapse issue of MAE, and propose U-MAE that enhances feature diversity with a uniformity loss. Empirically, U-MAE improves the linear probing accuracy of MAE by a large margin across different real-world datasets and backbones.

2 Related work

Self-Supervised Learning. Canonical deep learning relies on labeled data (*e.g.,* ImageNet) to train deep models. Instead, self-supervised learning (SSL) aims to learn meaningful representations from fully unlabeled data. A particular example is contrastive learning (CL) that learns to align augmented samples in the feature space, which achieves remarkable success and largely closes the performance gap between supervised and self-supervised learning [4, 14, 12, 27]. Recently, Masked Image Modeling (MIM), stemming from the Masked Language Modeling (MLM) paradigm widely adopted in NLP (*e.g.,* BERT [9]), also shows promising results in visual representation learning, to name a few, BEiT [1], iBOT [32], SimMIM [29], and MAE [15]. Among these MIM methods, the former three (similar to BERT) have close connections to contrastive learning (their objective can be directly formulated as contrastive loss [20]), while MAE has several key differences to BERT-like methods. As the name suggests, MAE is more like an autoencoder rather than a token predictor: it adopts a *pixel-level* reconstruction loss, omits the *masked tokens* in the encoder input, and utilizes a fully *non-linear* encoder-decoder architecture. In this work, we mainly provide a theoretical understanding of the working mechanism of MAE, in particular, how it learns generalizable features. Meanwhile, our analysis can also be applied to other MIM methods by different specifications of the encoder-decoder and their inputs.

Understanding MAE. Despite the impressive success of MAE, the theoretical understanding of it is largely underexplored. Among several concurrent attempts, Cao *et al.* [2] mostly focus on the attention mechanism of MAE through an integral kernel perspective. A similar theoretical analysis [23] also shows that autoencoders can preserve useful semantics in the pretraining data. However, we notice that these MAE analyses mostly focus on the autoencoder architecture, while they largely ignore the role of masking. As shown by He *et al.* [15], the patchwise masking strategy is a key component that distinguishes MAE from standard autoencoders, and different mask ratios ρ have a large impact on the downstream performance of pretrained features of MAE. Therefore, our work aims to explain how the masking-based reconstruction task learns meaningful representations.

Downstream Generalization of SSL. Motivated by the empirical success of SSL, many researchers try to theoretically understand how SSL works. Arora *et al.* [25] establish downstream guarantees of contrastive learning representations. Wang *et al.* [28] revise their bounds by resolving the class collision issues, and develop a new understanding of contrastive learning from the perspective of augmentation overlap. Haochen *et al.* [13] propose an augmentation graph framework and characterize the downstream performance of the eigen-decomposition solution. While these prior works focus on the contrastive learning method, there are little theoretical understanding on the downstream performance on MAE. In this work, we establish the first theoretical guarantee on the downstream performance among MAE methods and discuss the effect of mask ratios theoretically and empirically.

3 Masked Autoencoders Perform Implicit Contrastive Learning

In this section, we establish a formal connection between MAE and contrastive learning by showing that MAE implicitly aligns positive input pairs induced from the masking mechanism. Specifically, in Section 3.2, we show how a small MAE loss implies a small alignment loss. Motivated by this connection, in Section 3.3, we further study its feature collapse issue and point out that MAE suffers from dimensional feature collapse. To address this issue, in Section 3.4, we propose a Uniformity-enhanced (U-MAE) loss to further promote feature diversity.

3.1 Mathematical Formulation for MAE

We begin by introducing a mathematical formulation of MAE [15]. Given a natural image $\bar{x}$ from an unlabeled dataset $\mathcal{D}_u$, we first reshape it into n patches, denoted as $\bar{x} \in \mathbb{R}^{n \times s}$ where s denotes the patch size (*e.g.,* 16×16 in ViT [10]). Then, we draw a random binary mask $m \in \{0, 1\}^n$ (drawing 0 with probability ρ, *i.e.,* the mask ratio), and obtain two complementary masked views of $\bar{x}$:

$$x_1 = \bar{x}[m] \in \mathbb{R}^{n_1 \times s}, \quad x_2 = \bar{x}[1-m] \in \mathbb{R}^{n_2 \times s}, \tag{1}$$

where $n_1 = n(1-\rho), n_2 = n\rho$ are integers satisfying $n = n_1 + n_2$. We denote this random masking process as drawing x_1, x_2 from the joint distribution $\mathcal{M}(x_1, x_2|\bar{x})$, whose marginal distributions are $\mathcal{M}_1(x_1|\bar{x})$ and $\mathcal{M}_2(x_2|\bar{x})$, respectively. The MAE model $h = g \circ f$ is an encoder-decoder architecture, where an encoder f maps inputs x_1 to a latent feature $z_1 = f(x_1)$, and a decoder g maps the latent feature z_1 back to the pixel space to reconstruct the complementary target view x_2. Specifically, MAE adopts a simple mean square error (MSE) loss:

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

where the decoder output $\hat{x}_2 = h(x_1) = g(f(x_1))$ is assumed to be l_2-normalized following the original paper of MAE saying that normalized target x_2 yields better performance [15]. Although our analysis is based on MAE, it is quite general and can be naturally transferred to other Masked Image Modeling (MIM) frameworks, such as BEiT [1], iBOT [32], and SimMIM [29]. This is because their differences mainly lie in the implementation details, which is discussed in detail in Appendix C.

The Bipartite Mask Graph $\mathcal{G}_M$ of MAE. We notice that MAE essentially learns to pair the two complementary views x_1, x_2 via the reconstruction task, which can be modeled 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[3] (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.

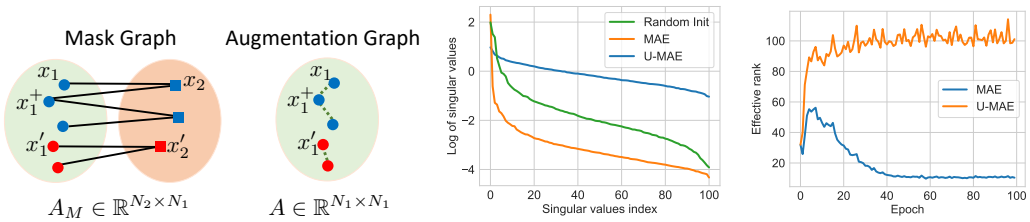
Considering the masking process in Eq. 1, there only exist edges *between* the two sets $\mathcal{X}_1, \mathcal{X}_2$ and there is no edge *within* each set itself. Thus, the mask graph $\mathcal{G}_M$ is a *bipartite* graph. Its adjacency matrix can be simply defined as $A_M \in \mathbb{R}^{N_2 \times N_1}$ 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.

3.2 MAE Implicitly Aligns Positive Input Pairs

As the name suggests, MAE is basically composed of two key designs: mask and autoencoders. For simplicity, we begin by assuming that the encoder-decoder architecture of MAE is capable of accomplishing the *vanilla* autoencoder task, *i.e.,* reconstructing the original input.

Assumption 3.1. *For any non-degenerate decoder g, we assume the existence of a pseudo-inverse encoder $f_g \in \mathcal{F}$ such that the resulting pseudo autoencoder $h_g = g \circ f_g$ satisfies $\mathbb{E}_x \|h_g(x) - x\|^2 \leq \varepsilon$, where x represents either unmasked data x_1 or masked data x_2.*

[3] This is used to avoid non-essential nuances related to functional analysis. Our discussion can also be extended to the infinite data regime following Haochen *et al.* [13].



(a) mask graph and augmentation graph (b) singular values (c) effective rank

Figure 1: (a) An illustration of the mask graph and the corresponding augmentation graph of MAE. Different colors denote different belonging classes. (b) Comparison of the singular values of learned features with 1) random initialization, 2) MAE loss, 3) our U-MAE loss. (c) The changing process of effective rank [24] of the encoded features trained with different objectives (MAE and U-MAE).

This assumption is likely to hold in practice, since deep networks already demonstrate excellent performance on the autoencoding task [19, 17]. Besides, some works show that the Transformer network adopted in MAE has universal approximation ability [30]. However, we note that vanilla autoencoder task cannot learn as meaningful representations as MAE. For example, without applying any masks, the linear probing accuracy of MAE drops dramatically from 61.2% to 17.4% on ImageNet-100. This indicates that the autoencoding ability (studied in [2, 23]) is *not* enough to explain the efficacy of MAE, which motivates us to further explore the masking mechanism in MAE.

MAE Performs Asymmetric Input-output Alignment on the Mask Graph. First, we show that the MAE loss can be lower bounded by an *asymmetric* alignment loss between the two complementary views x_1, x_2 using two-branch autoencoders h (real) and h_g (pseudo), respectively.

Theorem 3.2. *Under Assumption 3.1, the MAE loss can be lower bounded by*

$$\mathcal{L}_{MAE}(h) \geq \mathcal{L}_{asym}(h) - \varepsilon + const, \tag{3}$$

$$\text{and } \mathcal{L}_{asym}(h) = -\mathbb{E}_{x_1,x_2} h(x_1)^\top h_g(x_2) = -\operatorname{tr}(H^\top \bar{A}_M H_g), \tag{4}$$

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 $(H_g)_{x_2} = \sqrt{d_{x_2}} h_g(x_2)$.

As a result, a small MAE loss implies a small alignment loss (as a lower bound), which indicates that MAE will implicitly align the masked and unmasked views with an encoder-decoder architecture. However, it is a little bit confusing why aligning the two complementary views helps learn meaningful features. For a closer look, we find that via masking, MAE actually produces *implicit connections* among different *input samples* in the form of *2-hop connectivity*. Consider a pair of 2-hop input neighbors $x_1, x_1^+ \in \mathcal{X}_1$ that share a common complementary target view $x_2 \in \mathcal{X}_2$ (more likely to happen under a larger mask ratio ρ), as illustrated in Figure 1(a). By enforcing x_1, x_1^+ to reconstruct the same output x_2, MAE will implicitly map their features together. In this way, the 2-hop input neighbors serve as *positive pairs* that will be implicitly aligned as in contrastive learning.

Motivated by this observation, we seek to establish a formal connection between MAE and contrastive learning below. In particular, we will show that like the data augmentations in contrastive learning, the masking mechanism of MAE also implicitly introduces affinities between *input samples* and create (implicit) *positive pairs*. For a formal exposure, we define an augmentation graph $\mathcal{G}_A$ for modeling the relationship between all input samples in $\mathcal{X}_1$, which is different from the mask graph $\mathcal{G}_M$ for modeling the input-output relationship between $\mathcal{X}_1$ and $\mathcal{X}_2$.

The Augmentation Graph $\mathcal{G}_A$ of MAE. We construct an *augmentation graph* $\mathcal{G}_A$ for modeling the *mask-induced affinity* between all unmasked views in $\mathcal{X}_1$ [4]. Specifically, for any two views $x_1, x_1^+ \in \mathcal{X}_1$, we define their edge weight in the adjacency matrix A as the probability of having the same target view, *i.e.*, $\mathcal{A}(x_1, x_1') = \mathbb{E}_{x_2} \mathcal{M}(x_1|x_2)\mathcal{M}(x_1'|x_2)$ [5]. Note that the augmentation graph

[4] 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.

[5] $\mathcal{M}(x_1|x_2) = \mathcal{M}(x_1, x_2)/\mathcal{M}(x_2)$ can further be calculated by marginalizing $\mathcal{M}(x_1, x_2|\bar{x})$.

has the same degree matrix D_1 as the mask graph. Based on this formulation, we define a *symmetric* alignment loss *w.r.t.* the positive pairs $(x_1, x_1^+) \sim \mathcal{A}(x_1, x_1^+)$ drawn according to the augmentation graph:

$$\mathcal{L}_{\text{align}}(h) = -\mathbb{E}_{x_1, x_1^+} h(x_1)^\top h(x_1^+). \tag{5}$$

MAE Performs Symmetric Input Alignment on the Augmentation Graph. Built upon the augmentation graph constructed above, we theoretically verify the intuition on 2-hop connectivity by establishing the 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 3.3. *The asymmetric alignment loss on the mask graph (Eq. (4)) can be lower bounded by the symmetric alignment loss on the augmentation graph (Eq. (5)):*

$$\mathcal{L}_{asym}(h) \geq \frac{1}{2}\mathcal{L}_{align}(h) + const. \tag{6}$$

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 $\mathcal{L}_{asym}(h)$ with an arithmetic inequality, and then establish its equivalence to the symmetric alignment loss $\mathcal{L}_{align}(h)$. The derivation highlights the intrinsic connection between the mask graph (adjacency matrix A_M) and the augmentation graph (adjacency matrix A).*

$$\begin{aligned}
\mathcal{L}_{asym}(h) &= \mathbb{E}_{x_1, x_2} h(x_1)^\top h_g(x_2) \\
&= -\mathrm{tr}(H^\top \bar{A}_M H_g) && \textit{(reformulated to the mask graph (Theorem 3.2))} \\
&\geq -\frac{1}{2}(\|H^\top \bar{A}_M\|^2 + \|H_g\|^2) && \textit{(because } \mathrm{tr}(AB) \leq \frac{1}{2}(\|A\|^2 + \|B\|^2)\textit{)} \\
&= -\frac{1}{2}\mathrm{tr}(H^\top A_M A_M^\top H) - \frac{1}{2} && \textit{(because } \|H_g\|^2 = \sum_{x_2} d_{x_2}\|h_g(x_2)\|^2 = 1\textit{)} \\
&= -\frac{1}{2}\sum_{x_1, x_1'} A_{x_1, x_1'} h(x_1)^\top h(x_1') - \frac{1}{2} && \textit{(transformed to the augmentation graph)} \\
&= \frac{1}{2}\mathcal{L}_{align}(h) - \frac{1}{2}. && \textit{(following the definition in Eq. (5))}
\end{aligned}$$

Combining Theorem 3.2 and Theorem 3.3, we arrive at the main theorem showing that MAE'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.

Theorem 3.4. *Under Assumption 3.1, MAE's reconstruction loss (Eq. (2)) can be lower bounded by the alignment loss between positive pairs $(x_1, x_1^+) \sim \mathcal{A}(x_1, x_1^+)$,*

$$\mathcal{L}_{MAE}(h) \geq \frac{1}{2}\mathcal{L}_{align}(h) - \varepsilon + const = -\frac{1}{2}\mathbb{E}_{x_1, x_1^+} h(x_1)^\top h(x_1^+) - \varepsilon + const. \tag{7}$$

In this way, we establish a close relationship between the two mainstream SSL paradigms (MAE and contrastive learning) by showing that a small MAE 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 MAE (discussed in Section 4).

Comparison to Contrastive Learning. Comparing the alignment loss (Eq. (5)) to that of contrastive learning [22, 4, 13], we notice that the main difference lies in that contrastive learning aligns features in the *latent* space of the encoder f, while MAE aligns features in the *output* space with an encoder-decoder architecture $h = g \circ f$. Nevertheless, we also note that most variants of contrastive learning apply a nonlinear projection head g upon the encoder f before calculating the alignment loss [4, 5, 14, 12, 6], and MAE also adopts a shadow decoder g. Therefore, we may also regard the role of MAE'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 MAE loss is further lower bounded by an alignment loss defined in the feature space.

Corollary 3.5. *Under Assumption 3.1 and the assumption that the decoder is L-bi-Lipschitz, i.e., $\forall\,(x_1, x_2), 1/L\|x_1 - x_2\|^2 \leq \|g(x_1) - g(x_2)\|^2 \leq L\|x_1 - x_2\|^2$. Then, the MAE reconstruction loss can be lower bounded by the alignment loss w.r.t. the encoder outputs:*

$$\mathcal{L}_{MAE}(h) \geq -1/(2L) \cdot \mathbb{E}_{x_1, x_1^+} f(x_1)^\top f(x_1^+) - \varepsilon + const. \tag{8}$$

3.3 The Feature Collapse Issue in MAE

We have revealed that the MAE loss is closely related to an alignment loss. However, it is well known that in contrastive learning, simply aligning the positive pairs will result in the full feature collapse, because the alignment loss can also be minimized when the encoder produces a constant feature for all inputs. There, various techniques are proposed to address the issue by incorporating an additional loss to encourage feature uniformity or feature decorrelation [4, 14, 31], or by asymmetric structural designs [12, 6, 3]. MAE does not implement any of these techniques, but its latent features still do not fully collapse. One would wonder how MAE attains this property. In the following theorem, we show that minimizing the MAE loss can provably get rid of the full feature collapse.

Theorem 3.6. *When the encoder fully collapses, i.e., $\forall x \in \mathcal{X}_1, f(x) = c$, the MAE loss has a large lower bound:*

$$\mathcal{L}_{MAE}(h) \geq \mathrm{Var}(x_2), \tag{9}$$

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

MAE Can Avoid Full Feature Collapse. As the training data contain diverse images, the variance $\mathrm{Var}(x_2)$ will be relatively large. Therefore, unlike the alignment loss in contrastive learning, the MAE loss cannot be minimized (to a small value) by a collapsed encoder. The main reason is that alignment loss operates in a fully flexible latent space that allows a collapsed encoder to minimize the loss, while in MAE, the reconstruction loss adopts a parameter-invariant and sample-dependent target x_2. In this case, a fully collapsed encoder can no longer minimize the reconstruction loss *w.r.t.* x_2, which, in turn, means that MAE will not fully collapse.

MAE Still Suffers from Dimensional Collapse. Although MAE can avoid full feature collapse, it could still suffer from *dimensional feature collapse* where the learned features lie in a low dimensional subspace [16, 18], which also limits its representation power. We empirically examine this issue on ImageNet-100. Figure 1(b) shows that after learning, MAE's features indeed become more collapsed as there are fewer large singular values (indicating non-collapse dimensions). Quantitatively, Figure 1(c) shows that the features of MAE gradually collapse during the training process, as the effective rank [24] becomes smaller and smaller. This shows that MAE suffers from an increasing degree of dimensional feature collapse. Next, we introduce an explicit regularization to address this issue.

3.4 U-MAE: Enhancing Feature Diversity with an Explicit Uniformity Regularization

To further enhance the feature diversity of MAE, inspired by the uniformity loss in contrastive learning [22, 4, 13], we propose the following Uniformity-enhanced MAE (U-MAE) loss with an explicit regularization on feature uniformity through a coefficient $\lambda > 0$,

$$\mathcal{L}_{\text{U-MAE}}(h) = \mathcal{L}_{\text{MAE}}(h) + \lambda \cdot \mathcal{L}_{\text{unif}}(f), \tag{10}$$

$$\text{where } \mathcal{L}_{\text{unif}}(f) = \mathbb{E}_{x_1}\mathbb{E}_{x_1^-}(f(x_1)^\top f(x_1^-))^2, \tag{11}$$

and x_1^- denotes an independently drawn unmasked view from $\mathcal{X}_1$. Intuitively, the spectral uniformity loss $\mathcal{L}_{\text{unif}}(f)$ [13] 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 1(b) & 1(c) that U-MAE effectively addresses the dimensional feature collapse issue and improves the effective feature dimensionality by a large margin.

Theoretically, combined with Corollary 3.5, we can show that the U-MAE loss is lower-bounded by the spectral contrastive loss with a specific choice of λ.

Theorem 3.7. *Denote the spectral contrastive loss (SCL) from Haochen et al. [13] as*

$$\mathcal{L}_{SCL}(f) = 2\mathcal{L}_{align}(f) + \mathcal{L}_{unif}(f) = -2\mathbb{E}_{x_1,x_1^+} f(x_1)^\top f(x_1^+) + \mathbb{E}_{x_1,x_1'}(f(x_1)^\top f(x_1^-))^2. \tag{12}$$

Under Assumption 3.1 and the assumption that the decoder is L-bi-Lipschitz, when $\lambda = 1/(4L)$, the U-MAE loss can be lower bounded by the SCL loss:

$$\mathcal{L}_{U\text{-}MAE}(h) \geq \frac{1}{4L} \cdot \mathcal{L}_{SCL}(f) - \varepsilon + const. \tag{13}$$

As a result, minimizing the U-MAE loss will implicitly minimize the spectral contrastive loss among input views. Leveraging this inherent connection, in the next section, we further explain the important role of masking on the downstream generalization of MAE (as well as U-MAE).

4 Downstream Generalization of Masked Autoencoders

In Section 3, we establish a close connection between MAE and contrastive learning. Therefore, leveraging existing understandings of contrastive learning, this connection could provide new insights to understand how the masking strategy affects MAE's downstream generalization. In Section 4.1, we first provide theoretical bounds on downstream performance based on the mask-induced augmentation graph introduced in Section 3.2. Built upon these theoretical insights, in Section 4.2, we further investigate the effect of mask ratios and explain the optimal mask ratio of MAE.

4.1 Theoretical Guarantees on Downstream Classification

In this part, leveraging the connection established in Section 3, we characterize the downstream performance of U-MAE on the c-class linear classification task [25, 13], which 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. Arora *et al.* [25] empirically show that the mean classifier can obtain comparable performance to learnable linear heads.

Theorem 4.1. *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-MAE pretraining loss:*

$$\Pr(\bar{y} \neq p_f(\bar{x})) \leq c_1 L \cdot \mathcal{L}_{\text{U-MAE}}(h) + c_2\alpha + c_3 L\varepsilon + c_4, \tag{14}$$

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

This theorem provides an upper bound on the downstream error of an encoder f with its U-MAE pretraining loss, which is the *first theoretical guarantee* on the downstream performance of MAE methods. As an implication of this theorem, a small U-MAE loss would provably imply a small downstream classification error, which helps explain the good downstream generalization ability of MAE [15]. Besides, we establish a common lower bound on the U-MAE loss that holds for all $h \in \mathcal{H}$. As a large common lower bound of U-MAE 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 4.2. *The U-MAE pretraining loss has the following common lower bound:*

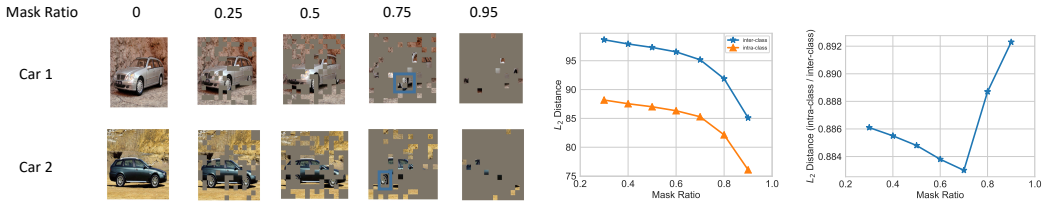
$$\forall\, h \in \mathcal{H}, \quad \mathcal{L}_{\text{U-MAE}}(h) \geq \frac{1}{4L} \sum_{i=k+1}^{N_1} \lambda_i^2 - \varepsilon + \textit{const}, \tag{15}$$

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

Combining Theorem 4.1 and Theorem 4.2, we can see that the downstream error of MAE can be minimized with a small vanilla autoencoding error ε, 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 ε depends on the capacity of the chosen model class $\mathcal{H}$, while the label error α and residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$ purely depend on the mask-induced augmentation graph $\mathcal{G}_A$. Therefore, overall speaking, MAE with large capacity can have a smaller downstream error with a smaller autoencoding error ε. In the meantime, the masking ratio ρ should be properly chosen to ensure a small label error α and small magnitude of residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$. Indeed, He *et al.* [15] show that the mask ratio has a decisive influence on the downstream performance of MAE. Therefore, in the next part, we discuss how the choice of mask ratio would affect the downstream generalization of MAE via the label error α and residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$.

4.2 The Effect of Mask Ratio on Downstream Generalization

A surprising fact in MAE is that the optimal mask ratio is pretty high, *e.g.,* $\rho = 0.75$, where most image contents are lost. Therefore, we are wondering why such a high mask ratio is necessary and how it helps (instead of damaging) the learning of data representations. In this part, we provide both theoretical and empirical insights on this problem based on our proposed theory.



(a) Masked views with different mask ratio

(b) The distance between intra-class and inter-class samples

(c) The relative distance between intra-class samples and inter-class samples

Figure 2: (a) Appropriate mask ratio can generate similar views from different samples in the same class. (b) The influence of mask ratio on the average l_2 distance on ImageNet-100 between intra-class and inter-class samples. (c) The influence of mask ratio on the relative distance between intra-class and inter-class samples.

Theoretical Insights. As discussed above, Theorems 4.1 & 4.2 suggest that a small label error α and small magnitude of residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$ are important for good downstream performance in MAE. In particular, we notice that the mask ratio ρ has a decisive influence on both factors. On the one hand, the label error α will grow with a larger mask ratio, but it is only very large under an extremely high mask ratio. Intuitively, α represents the possibility of recovering the original class y from the left patches after masking. As illustrated in Figure 2(a), small or medium mask ratio, even a large one of 0.75, hardly alters the belonging class, and only when the mask ratio is extremely high, *e.g.,* 0.95, we can hardly tell it is still a car. On the other hand, according to the spectral graph theory [7], the residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$ represent the high frequency components of the graph, which have large magnitude when the graph is less connected, *e.g.,* having many disjoint components. Therefore, a small downstream error requires the augmentation graph $\mathcal{G}_A$ to have better connectivity. In this aspect, Figure 2(a) shows that when the mask ratio increases, many dissimilar patterns are masked out and the left patches can have an increasing level of similarity, particularly among intra-class samples, *e.g.,* the tires of the two cars. Therefore, a large mask ratio can effectively improve the connectivity of the augmentation graph by reducing sample variation and increasing inter-sample similarity. Considering both the effects on the label error α and the residual components $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$, we notice that there is a tradeoff in the choice of mask ratio: we should choose a properly large one to increase graph connectivity, and avoid too large mask ratio as it leads to the class mixture. In other words, a guiding principle is that a mask ratio should be chosen such that the mask-induced graph connectivity should happen mostly among *intra-class* samples (no harm) instead of *inter-class* samples (inducing large label error). Below, we further examine this understanding on real-world data.

Empirical Investigation. To verify our perspective, we conduct experiments on CIFAR-10 with different mask ratios. We evaluate the distance between two images by calculating the average l_2 distance between every pair of patches. As shown in Figure 2(b), we find that the average distance of intra-class images decreases with the increasing of mask ratio, which means that a higher mask ratio leads to generating more edges between intra-class samples with better connectivity (*i.e.,* decreasing magnitude of residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$). Meanwhile, we also notice that a high mask ratio will also lead to a smaller inter-class distance, which suggests an increasing label error α. Therefore, there exists a tradeoff on the mask ratio to balance residual eigenvalues $\{\lambda_{k+1}, \ldots, \lambda_{N_1}\}$ and labeling error α.

The Sweet Spot for Mask Ratio. The discussion above reveals that mask ratio has an effect on both intra-class and inter-class connectivity, and a good mask ratio should be selected to balance the two sides such that intra-class distance is relatively high while inter-class distance is relatively low. Motivated by this, we compute the relative distance (intra-class over inter-class). As shown in Figure 2(c), the relative distance decreases first with the increasing mask ratio, showing that the intra-class distance decreases faster than the inter-class distance under small mask ratio. When $\rho > 0.7$, the relative distance becomes larger again, indicating that the difference between intra-class and inter-class edges disappears under too large mask ratio. The sweet spot lies in $\rho = 0.7$, which is pretty close to the optimal mask ratio of MAE ($\rho = 0.75$). This shows that our theoretical analysis of the effect of mask ratio aligns surprisingly well with the practice of MAE.

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

		CIFAR-10		ImageNet-100		ImageNet-1K	
Downstream Task	Method	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-MAE	**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-MAE	89.4	**90.8**	**86.8**	**87.3**	**83.0**	83.2

5 Experiments

In this section, we first present the main empirical results of our proposed U-MAE loss on different real-world datasets with different backbones. Then we conduct a series of experiments to understand how well the U-MAE loss works.

5.1 Evaluation on Benchmark Datasets

To evaluate the effectiveness of the proposed U-MAE loss, extensive experiments are conducted on CIFAR-10 [21], ImageNet-100 [8], and ImageNet-1K [8].

Setup. We mainly follow the basic setup of MAE [15]: for the encoder, we adopt different variants of ViT [10], *i.e.,* ViT-Tiny, ViT-Base, and ViT-Large. For the decoder, we use a flexible one following [15]. The mask ratio is set to 0.75. For U-MAE, the coefficient of the uniformity term is set to 0.001. 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. We conduct both linear evaluation and non-linear fine-tuning on the pretrained encoder. For linear evaluation, we train a linear classifier on the frozen pretrained encoder. As for non-linear fine-tuning, we train both the pretrained encoder and the linear classifier with the cross entropy loss.

Effectiveness of the Proposed U-MAE Loss. In Table 1, we compare the performance of original MAE loss and U-MAE loss on different benchmarks. We find that, on linear evaluation results, our proposed U-MAE loss increases 8.9% on CIFAR-10, 7.2 % on ImageNet-100, and 3.4% on ImageNet-1K with two different backbones. On fine-tuning results, our proposed U-MAE loss will not hurt the performance of fine-tuning results of MAE. The experimental results verify the effectiveness of the proposed U-MAE loss, which achieves better performance than the original MAE loss across different datasets and different backbones. Moreover, to quickly evaluate the training process of the encoder, we set an online linear classifier to monitor the linear accuracy whose results can be found in Appendix B.5.

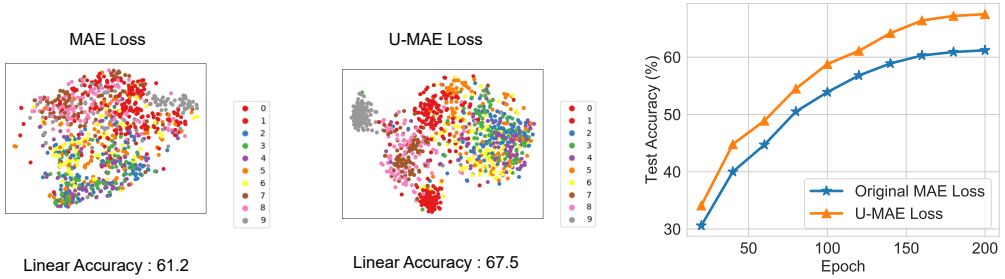
Extention to Other MIM Frameworks. We also verify our proposed method in another MIM framework SimMIM [29]. For SimMIM, we use ViT-Base as the encoder and use the linear decoder as in [29]. We use the recommended mask ratio 0.6. Similar to U-MAE, Uniformity-enhanced SimMIM loss (U-SimMIM) adds a uniformity regularization term to the original reconstruction loss of SimMIM, where the coefficient of the uniformity term is set to 0.0001. 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 2. We can see that the linear accuracy increases 6.8% with the uniformity regularizer, which further verifies the general property of our approach.

Table 2: Linear probing accuracy (%) of U-SimMIM (ViT-Base) on ImageNet-100.

SimMIM	U-SimMIM
54.3	**61.1**

Table 3: Linear probing accuracy (%) of U-MAE with different coefficients (Eq. (10)).

λ	0	1e-5	1e-4	1e-3	1e-2
Acc	61.2	65.9	**67.5**	45.1	47.0



(a) t-SNE representations

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

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

5.2 Empirical Understandings

Visualization of Representations. To intuitively understand the improvement of our U-MAE loss on clustering intra-class samples, we use t-SNE [26] to visualize the representations trained with MAE loss and U-MAE loss on ten random class of ImageNet-100 (detailed classes are introduced in Appendix B.4). 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 be distinguished, especially with our uniformity regularizer.

Different Coefficients of the Regularizer Term. The most important hyper-parameter of our proposed U-MAE loss is the coefficient of the uniformity regularizer term. In Table 3, we present the results of linear evaluation on ImageNet-100 trained with the U-MAE loss with different coefficients of the regularizer term. We can see that the downstream performance increases when the coefficient increases from 0 to 0.0001. However, the overlarge coefficient will also hurt the performance of U-MAE loss as the task of MAE will be overlooked.

Training process. To further compare the performance between the original MAE loss and U-MAE 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-MAE loss improves the performance of MAE with all different training epochs, which verifies the effectiveness of our proposed U-MAE loss.

6 Conclusion

In this paper, we proposed a new theoretical understanding of MAE by formally connecting MAE and contrastive learning. In particular, we show that a small MAE loss implies a good alignment between the mask-induced positive samples. Based on this connection, we further analyze the dimensional collapse issue of MAE and propose a new variant, named Uniformity-enhanced MAE (U-MAE), through adding an explicit feature uniformity regularization. Theoretically, this connection enables us to establish theoretical guarantees on the downstream classification error, which also provides theoretical insights on the choice of large mask ratio in MAE. Empirically, the proposed U-MAE successfully mitigates the dimensional collapse issue of MAE, and achieves consistent improvement on the linear probing task on CIFAR-10, ImgaeNet-100, and ImageNet-1K.

Acknowledgment

Yisen Wang is partially supported by the NSF China (No. 62006153), Project 2020BD006 supported by PKU-Baidu Fund, and Huawei Technologies Inc. We thank anonymous reviewers for constructive and helpful discussions.

References

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

[2] Shuhao Cao, Peng Xu, and David A. Clifton. How to understand masked autoencoders. *CoRR*, 2022.

[3] 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.

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

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

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

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

[8] 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.

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

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

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

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

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

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

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

[17] Li Jing, Jure Zbontar, et al. Implicit rank-minimizing autoencoder. In *NeurIPS*, 2020.

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

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

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

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

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

[23] Jiachun Pan, Pan Zhou, and Shuicheng Yan. Towards understanding why mask-reconstruction pretraining helps in downstream tasks. *arXiv preprint arXiv:2206.03826*, 2022.

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

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

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

[27] 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.

[28] 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.

[29] 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.

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

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

A.1 Proof of Theorem 3.4

Proof. With Assumption 3.1, we have

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

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 column of the matrix H and $H_g(x_2)=\sqrt{d_{x_2}}h(x_2)$ 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,

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

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

$$\mathcal{L}_{\text{MAE}}(h) \geq -\frac{1}{2}(\operatorname{tr}(\bar{A}_M^\top \bar{A}_M H H^\top) + 1) - \varepsilon + 1 = -\frac{1}{2}\operatorname{tr}(\bar{A}_M^\top \bar{A}_M H H^\top) - \varepsilon + \frac{1}{2}. \tag{16}$$

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

$$(\bar{A}_M^\top \bar{A}_M)_{x_1,x_1^+} = \sum_{x_2} \frac{w_{x_1,x_2} w_{x_1,x_2^+}}{d_{x_2}\sqrt{d_{x_1}d_{x_1^+}}}. \tag{17}$$

$$(HH^\top)_{x_1^+,x_1} = \sqrt{d_{x_1}d_{\ x_1^+}}h(x_1)^\top h(x_1^+). \tag{18}$$

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.,*

$$(\bar{A}_M^\top \bar{A}_M H H^\top)_{x_1,x_1} = \sum_{x_1^+}(A_M^\top A_M)_{x_1,x_1^+}(HH^\top)_{x_1^+,x_1} = \sum_{x_1^+}\sum_{x_2}\frac{w_{x_1,x_2}w_{x_1^+,x_2}}{d_{x_2}}h(x_1)^\top h(x_1^+). \tag{19}$$

With that, we can element-wise expand Eq. (16),

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

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

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

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 3.5

With Theorem 3.4, 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

$$\forall\,(x_1,x_2), 1/L\cdot\|x_1-x_2\|^2 \leq \|g(x_1)-g(x_2)\|^2 \leq L\cdot\|x_1-x_2\|^2. \tag{20}$$

So,

$$
\begin{aligned}
\mathcal{L}_{\text{MAE}}(h) &\geq -\frac{1}{2}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} h(x_1)^\top h(x_1^+) - \varepsilon + \frac{1}{2} \\
&= \frac{1}{4}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} \|h(x_1) - h(x_1^+)\|^2 - \varepsilon && (h(x) \text{ is normalized}) \\
&= \frac{1}{4}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} \|g(f(x_1)) - g(f(x_1^+))\|^2 - \varepsilon \\
&\geq \frac{1}{4L}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} \|f(x_1) - (f(x_1^+)\|^2 - \varepsilon && (\text{Equation (20)}) \\
&= -\frac{1}{2L}\mathbb{E}_{x_1,x_1^+\sim\hat{p}(x_1,x_1^+)} f(x_1)^\top f(x_1^+) - \varepsilon + \frac{1}{2}.
\end{aligned}
$$

A.3 Proof of Theorem 3.6

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

$$\forall x \in \mathcal{X}_1, f(x) = c. \tag{21}$$

Then,

$$
\begin{aligned}
\mathcal{L}_{MAE}(h) &= \mathbb{E}_{x_1,x_2}\|g(f(x_1)) - x_2\|^2 \\
&= \mathbb{E}_{x_2}\|g(c) - x_2\|^2
\end{aligned} \tag{22}
$$

We then select a g(c) to make Equation (22) minimal. According to KKT conditions, it has a closed-form solution $q^\star$, satisfying

$$2\mathbb{E}_{x_2}(q^\star - x_2) = 0. \tag{23}$$

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

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

□

A.4 Proof of Theorem 3.7

With Corollary 3.5, we have

$$\mathcal{L}_{\text{MAE}}(h) \geq -1/(2L) \cdot \mathbb{E}_{x_1,x_1^+} f(x_1)^\top f_g(x_1^+) - \varepsilon + const. \tag{24}$$

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

$$
\begin{aligned}
\mathcal{L}_{\text{U-MAE}}(h) &= \mathcal{L}_{MAE}(h) + 1/(4L) \cdot \mathcal{L}_{unif}(f) \\
&\geq 1/(2L) \cdot \mathcal{L}_{align}(f) + 1/(4L) \cdot \mathcal{L}_{unif}(f) - \varepsilon + const \\
&= 1/(4L) \cdot \mathcal{L}_{SCL}(f) - \varepsilon + const.
\end{aligned} \tag{25}
$$

A.5 Proof of Theorem 4.1

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.,* $A_{x_1,x_1^+} = \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1} \cdot d_{x_1^+}}}$. Then we reformulate the downstream error,

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

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 $(D^{1/2}Y - \bar{A}C)$,

$$(D^{1/2}Y)_{x_1,j} = \sqrt{(d_{x_1})\mathbb{1}_{y_{x_1}=j}}, \quad (\bar{A}C)_{x_1,j} = \sum_{x_1^+} \frac{w_{x_1,x_1^+}}{\sqrt{d_{x_1}} \cdot \sqrt{d_{x_1^+}}} \sqrt{(d_{x_1^+})\mathbb{1}_{y_{x_1^+}=j}}. \tag{26}$$

So when $j = y_{x_1}$,

$$\begin{aligned}
(D^{1/2}Y - \bar{A}C)_{x_1,j} &= \sqrt{(d_{x_1})\mathbb{1}_{y_{x_1}=j}} - \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\
&= \sqrt{d_{x_1}} - \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\
&= \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}} - \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\
&= \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq j} \\
&= \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}}.
\end{aligned} \tag{27}$$

When $j \neq y_{x_1}$,

$$\begin{aligned}
(D^{1/2}Y - \bar{A}C)_{x_1,j} &= \sqrt{(d_{x_1})\mathbb{1}_{y_{x_1}=j}} - \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\
&= 0 - \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j} \\
&= -\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j}.
\end{aligned} \tag{28}$$

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

$$\begin{aligned}
\|(D^{1/2}Y - \bar{A}C)_{x_1}\|^2 &= (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}})^2 + \sum_{j\neq y_{x_1}} (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j})^2 \\
&\leq (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}})^2 + (\sum_{j\neq y_{x_1}} \sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}=j})^2 \\
&\leq (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}})^2 + (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}} \sum_{j\neq y_{x_1}} \mathbb{1}_{y_{x_1^+}=j})^2 \\
&= (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}})^2 + (\sum_{x_1^+} \frac{\mathcal{A}(x_1,x_1^+)}{\sqrt{d_{x_1}}}\mathbb{1}_{y_{x_1^+}\neq y_{x_1}})^2 \\
&= \frac{2\beta_{x_1}^2}{d_{x_1}}.
\end{aligned} \tag{29}$$

With that, we obtain $\|D^{1/2}Y - \bar{A}C\| = \sum_{x_1} \frac{2\beta_{x_1}^2}{d_{x_1}}$. As we assume that $E_{\bar{x}\sim\mathcal{P}_d}(\mathcal{A}(x_1|\bar{x})\mathbb{1}_{y_{x_1}\neq\bar{y}}) \leq \alpha$, so

$$\begin{aligned}
\sum_{x_1,x_1^+} \mathcal{A}(x_1,x_1^+)\mathbb{1}_{y_{x_1^+}\neq y_{x_1}} &= \sum_{x_1,x_1^+} E_{\bar{x}}(\mathcal{A}(x_1|\bar{x})\mathcal{A}(x_{x_1^+}|\bar{x})\mathbb{1}_{y_{x_1^+}\neq y_{x_1}}) \\
&\leq \sum_{x_1,x_1^+} E_{\bar{x}}(\mathcal{A}(x_1|\bar{x})\mathcal{A}(x_{x_1^+}|\bar{x})(\mathbb{1}_{y_i\neq\bar{y}} + \mathbb{1}_{y_{x_1^+}\neq\bar{y}})) \\
&= 2E_{\bar{x}\sim\mathcal{P}_d}(\mathcal{A}(x_1|\bar{x})\mathbb{1}_{y_{x_1}\neq\bar{y}}) \\
&\leq 2\alpha.
\end{aligned} \tag{30}$$

Then we have

$$\begin{aligned}
\|D^{1/2}Y - \bar{A}C\| &= \sum_{x_1} \frac{2\beta_{x_1}^2}{d_{x_1}} \\
&\leq \sum_{x_1} 2\beta_{x_1} && (d_{x_1} = \sum_{x_1^+}\mathcal{A}(x_1,x_1^+) \geq \sum_{x_1^+}\mathcal{A}(x_1,x_1^+)\mathbb{1}_{y_{x_1}\neq y_{x_1^+}}) \\
&= 2\sum_{x_1,x_1^+} \mathcal{A}(x_1,x_1^+)\mathbb{1}_{y_{x_1}\neq y_{x_1^+}} && \text{(definition of } \beta_{x_1}) \\
&\leq 4\alpha. && \text{(Equation (30))}
\end{aligned}$$

Then we obtain

$$\begin{aligned}
&\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 \\
&= \|D^{1/2}Y - \bar{A}C + \bar{A}C - UW_f\|^2 \\
&\leq 2\|\bar{A}C - UW_f\|^2 + 8\alpha && (\|A+B\|^2 \leq \|A\|^2 + \|B^2\|) \\
&= 2\|(\bar{A} - UU^T + UU^T)C - UW_f)\|^2 + 8\alpha \\
&= 2\|(\bar{A} - UU^T)C + U(U^TC - W_f)\|^2 + 8\alpha \\
&\leq 4(\|(\bar{A} - UU^T)C\|^2 + \|U(U^TC - W_f)\|)^2 + 8\alpha && (\|A+B\|^2 \leq 2(\|A\|^2 + \|B\|^2)) \\
&\leq 4(\|(\bar{A} - UU^T)\|^2\|C\|^2 + \|U\|^2\|(U^TC - W_f)\|^2) + 8\alpha && (\|AB\| \leq \|A\|\|B\|) \\
&\leq 4(\|(\bar{A} - UU^T)\|^2 + \|(U^TC - W_f)\|^2) + 8\alpha && (\|C\| = 1, \|U\| = 1) \\
&= 4\mathcal{L}_{SCL}(f) + const + \|(U^TC - W_f)\|^2 + 8\alpha && \text{(Lemma B.8 in [13])} \\
&\leq 16L \cdot \mathcal{L}_{\text{U-MAE}}(h) + 8\mathbb{E}_y[\mathbb{E}_{x|y}f(x) - b_y]^2 + 8\alpha + 16L\varepsilon + const. && \text{(Theorem 3.7)} \\
&\leq 16L \cdot \mathcal{L}_{\text{U-MAE}}(h) + 8\alpha + 16L\varepsilon + const. && (W_f \text{ is a mean classifier})
\end{aligned} \tag{31}$$

In the next step, we analyze the prediction error. We denote $\bar{y}$ as the ground-truth label of original data $\bar{x}$. We first define a 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, $\bar{y} \neq c'_f(\bar{x})$ only happens when more than half of the views predict wrong labels. So

$$\begin{aligned}
&\Pr(\bar{y} \neq c'_f(\bar{x})) \\
&\leq 2\Pr(\bar{y} \neq c_f(x)) \\
&\leq 4\mathbb{E}_{\bar{x}\sim\mathcal{P}_d(x),x\sim\mathcal{M}_1(x|\bar{x})}\|\bar{y} - W_f f(x)\|^2 && \text{(Claim B.9 in [13])} \\
&\leq 8(\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 + \mathbb{E}_{\bar{x}\sim\mathcal{P}_d(x),x\sim\mathcal{M}_1(x|\bar{x})}\|y - \bar{y}\|^2) && (\|A+B\|^2 \leq \|A\|^2 + \|B^2\|) \\
&\leq 8(\mathbb{E}_{x,y}\|y - W_f f(x)\|^2 + 2\alpha) && \text{(definition of } \alpha) \\
&\leq 32\mathcal{L}_{SCL}(f) + 64\alpha + 16\alpha + const \\
&\leq 128L \cdot \mathcal{L}_{\text{U-MAE}}(h) + 64\alpha + 16\alpha + 128L\varepsilon + const \\
&\leq 128L \cdot \mathcal{L}_{\text{U-MAE}}(h) + 80\alpha + 128L\varepsilon + const.
\end{aligned}$$

□

A.6 Proof of Theorem 4.2

Proof. With Equation (31), we have

$$\mathcal{L}_{\text{U-MAE}}(h) \geq \frac{1}{4L}\|A - UU^\top\|^2 - \varepsilon + const. \tag{32}$$

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 [11], we obtain

$$\|(\bar{A} - U^\star(U^\star)^T)\| = \sum_{i=k+1}^{N_1} \lambda_i^2, \tag{33}$$

where $\lambda_{k+1} \cdots \lambda_{N_1}$ are the $N_1 - k$ largest eigenvalues of matrix $\bar{A}$. We denote $h^\star$ is the minimizer of $\mathcal{L}_{\text{U-MAE}}$ and $f^\star$ is 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

$$\begin{aligned}
\mathcal{L}_{\text{U-MAE}}(h) \geq \mathcal{L}_{\text{U-MAE}}(h^\star) &\geq \frac{1}{4L}\|A - U_{f^\star}(U_{f\star})^\top\|^2 - \varepsilon + const \\
&\geq \frac{1}{4L}\|A - U^\star(U^\star)^\top\|^2 - \varepsilon + const \\
&= \frac{1}{4L}(\sum_{i=k+1}^{N_1} \lambda_i)^2 - \varepsilon + const.
\end{aligned} \tag{34}$$

□

B Additional Experiment

B.1 Experiment Details of Computing Effective Rank

We conduct experiments on ImageNet-100 and use ViT-Base as the backbone. We compare two kinds of pretrained encoders: 1) the encoder trained with original MAE loss, 2) the encoder trained with U-MAE loss ($\lambda = 0.0001$). Then we store the normalized encoded features. We construct a feature matrix A with n (n=100) random samples and compute its singular values $\sigma_1, \ldots, \sigma_n$. Then, we compute the effective rank [24] of the feature matrix as follows. We first compute the distribution of sigular values, *i.e.,* $p_k = \frac{\sigma_k}{\|\sigma\|_1}$. Then we can obtain the effective rank with that: $Erank(A) = \exp(-\sum_{k=1}^{n} p_k \log(p_k))$.

B.2 Experiment Details of Section 4.2

We conduct the verification experiment of Section 4.2 on ImageNet-100. We set patch size to 16 and use random masking. When computing the distance, we compute the max l_2 distance between the patches of two images and denote it as the distance between two images. Then we compute the average distance of different images. When computing the distance of intra-class samples, we only compute the distance between the samples in the same class. While for inter-class distance, we only compute the distance between the samples of different classes. To reduce the amount of calculation, we random select 10 classes of ImageNet-100 for this experiment.

B.3 Additional Experiment Details of Section 5.1

For CIFAR-10, we train ViT-Tiny on 1×NVIDIA V100 with 1600 epochs, which needs about 23 hours, and ViT-Base on 1×NVIDIA V100 with 1600 epochs, which needs about 25 hours. For ImageNet-100, we train ViT-Base on 4×NVIDIA V100 with 200 epochs, which needs about 37 hours, and ViT-Large on 4×NVIDIA V100 with 200 epochs, which needs about 42 hours.

Table 4: Online linear accuracy of MAE and U-MAE loss.

	CIFAR-10		ImageNet-100		ImageNet-1K	
Method	ViT-Tiny	ViT-Base	ViT-Base	ViT-Large	ViT-Base	ViT-Large
MAE	52.9	59.5	37.5	39.5	39.7	43.4
U-MAE	**69.4**	**72.0**	**56.3**	**61.4**	**46.5**	**52.1**

Table 5: Online linear accuracy of SimMIM and U-SimMIM loss on ImageNet-100 with ViT-Base.

Method	Top-1 Accuracy (%)
SimMIM	21.59%
U-SimMIM	**30.05%(+8.46%)**

B.4 Details of Visualization Results in Figure 3(a)

We use t-SNE to to visualize the representations learned with MAE and U-MAE loss on random 10 classes of ImageNet. The ten classes are 0) "cock", 1) "hen", 2) "tiger shark, Galeocerdo cuvieri", 3) "tench, Tinca tinca", 4) "goldfish, Carassius auratus", 5) "hammerhead, hammerhead shark", 6) "electric ray, crampfish, numbfish, torpedo", 7) "stingray", 8) "great white shark, white shark, man-eater, man-eating shark, Carcharodon carcharias", 9) "indigo bunting, indigo finch, indigo bird, Passerina cyanea",

B.5 Online Linear Evaluation Experiments

Besides the offline linear evaluation results in Section 5, we also present the results obtained by an online linear classifier. Specifically, we train a linear head along the MAE training process and detach its gradients. From Table 4, we find that our promoting loss increases 14.71% for linear evaluation results on CIFAR-10 with two different backbones and increases 7.35 % on ImageNet-100 with two different backbones. And we could see that our U-MAE also significantly outperforms MAE on large-scale datasets by improving 6.77% Top-1 accuracy on ImageNet-1K.

B.6 Additional Experiments on SimMIM

For SimMIM, we use ViT-Base as the encoder and use the linear decoder as proposed in SimMIM [29]. We use the recommended mask ratio, 60%. Similar to MAE, we propose a Uniformity-promoting SimMIM loss (U-SimMIM) which adds a uniformity regularizer term to the original reconstruction loss of SimMIM. For the uniformity term of our proposed loss, we set the coefficient of the uniformity term to 0.01. For ImageNet-100, we pretrain the model for 200 epochs with batch size 128 and weight decay 0.05. We conduct linear evaluation on the unsupervised pretrained encoder. From Table 5, we could see that our U-SimMIM also significantly outperforms SimMIM on large-scale datasets by improving 8.46% Top-1 accuracy on ImageNet-1K.

C Extension to Other MIM Methods

The basic paradigm of current MIM frameworks is to reconstruct the masked patches from unmasked ones, while their differences mainly exist in the implementation details [1, 32, 29, 15], *e.g.,* 1) the input of encoder: including masked patches (SimMIM, iBOT, BEiT) or not (MAE); 2) the decoder: one-layer (SimMIM, iBOT) or multi-layer (MAE, BEiT); and 3) the target: RGB (SimMIM, MAE) or tokenized discrete tokens (iBOT, BEiT). In fact, our theoretical framework is quite general, and can be easily extended to these variants:

- **Input of Encoder.** If masked patches are adopted, the input of encoder could be modeled as $\tilde{x}_1 = (x_1, p_{x_1+x_2})$ and $\tilde{x}_2 = (x_2, p_{x_1+x_2})$, where p represents the position embeddings. In this case, both of them have the access to the entire position encoding $p_{x_1+x_2}$.
- **Decoder.** One-layer decoder is a special case of our multi-layer formulation. Moreover, the one-layer decoder obtains the invertibility and our analysis can be simplified.

- **Target.** The tokenized target only corresponds to a specific format of x_2, which does not affect our analysis of their alignment property.