How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders

Anonymous Author(s)
Affiliation
Address
email

Abstract

Masked AutoEncoder (MAE) based on a reconstruction task has risen to be a promising paradigm for self-supervised learning and achieves state-of-the-art performance across different benchmark datasets. However, despite its impressive empirical success, a theoretical analysis of it is still limited. In this paper, we propose a new theoretical understanding of how MAE works and why the choice of mask ratio is so important for MAE from a graph perspective. Based on the analysis of the MAE loss, we prove that the MAE loss can be upper bounded by an implicit alignment loss and propose an insight that MAE bridges different samples in the same class with an aggressive mask ratio. Then, we establish a guarantee for the downstream performance of MAE and analyze the trade-off on choosing the mask ratio. Motivated by our theory, we propose a Uniformity-promoting MAE (U-MAE) loss and find it can significantly improve the downstream performance of MAE on real-world datasets, including CIFAR-10 and ImageNet-100.

1 Introduction

Recently, self-supervised learning has been proposed as an effective framework to learn meaningful representations without access to labels. There are lots of works raising different surrogate objectives to improve the performance of the learned representation on downstream tasks. Among them, masked autoencoding defines self-supervised signals by masking some patches of samples and learns a surrogate loss based on the reconstruction task of the missing message. Bert [6] is a popular self-supervised framework based on masked autoencoding in natural language process and achieves impressive performance. Recently, a paradigm of masked autoencoding named Masked Autoencoder (MAE) [11] obtains empirical breakthroughs on various computer vision tasks and achieves state-of-the-art performance. However, despite its promising empirical success, a theoretical understanding of why MAE can learn representations that can improve the performance of downstream tasks is still limited.

The basic idea of MAE is quite simple, which masks several patches of an image and encodes the unmasked patches with the Vision Transformer (ViT) [7], then tries to reconstruct the missing patches with a decoder via a training objective of mean square error. Intuitively, the reconstruction task is an instance-level task instead of a class-level task. However, the empirical success of MAE claims that the reconstruction target can learn helpful representations for downstream classification tasks. So how to understand the training process of MAE can bridge the samples in the same class is quite important.

[2] is the first theoretical work to understand MAE, which explains MAE from a perspective of an integral kernel. However, there still exists a vacancy in their analysis. For example, their work doesn't analyze the core component of MAE, *i.e.,* the masking technique. In this paper, we try to further understand the training process of MAE and its components. We prove that MAE loss is

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

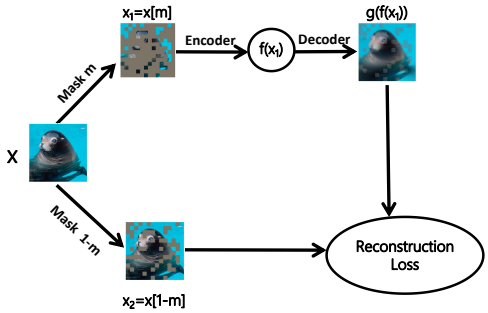


Figure 1: For each input x, Masked Autoencoder (MAE) first draws a random patchwise mask m, and obtain two complementary views x_1, x_2. Afterward, MAE utilizes an encoder-decoder architecture to reconstruct x_2 from x_1.

upper bounded by an implicit alignment loss and the reconstruction target aligns the features of the masked view and the unmasked view of the same image. Intuitively, some masked views of different images in the same class can be very similar, which means different images can be bridged by these similar overlapped views.

With the analysis of the MAE loss, we propose a guarantee for MAE training process and it will converge with some practical assumptions. Based on our analysis of MAE training process, we propose a Uniformity-promoting MAE (U-MAE) loss, which adds a regularizer term to the original MAE loss. Empirically, we find our proposed loss can significantly improve the performance of MAE on different benchmark datasets (CIFAR-10, ImageNet-100) with different backbones (ViT-Base, ViT-Large). We summarize our contributions as follows:

- We prove that MAE loss is upper bounded by an implicit alignment loss and develop a new understanding of how MAE bridges the samples in the same class. Based on our analysis, we propose a new U-MAE loss.
- We establish a guarantee for the downstream performance of MAE with practical assumptions and compare the training process of contrastive learning and MAE.
- We empirically verify that our proposed U-MAE loss can significantly improve the performance of MAE across different real-world datasets, including CIFAR-10 and ImaegNet-100.

2 Related work

Self-Supervised Learning. With meaningful surrogate objectives, self-supervised learning has gradually closed the performance gap between supervised and unsupervised learning without access to expensive supervised information. For example, contrastive learning designs appropriate data augmentations to draw together the semantically similar samples and obtain impressive downstream performance on various datasets [3, 10, 8]. Masked Image Modeling (MIM) is another surrogate task to learn meaning representations, which masks some patches of samples and trains the encoder based on the comparison between the masked and unmasked patches. Masked Autoencoder (MAE) [11] is a representative work of MIM and obtains state-of-the-art performance on various large-scale datasets like ImageNet. The core task of MAE is to reconstruct masked images with an asymmetric Vision Transformer architecture

The Theory of Self-supervised Learning. Inspired by the empirical success of self-supervised learning, many researches try to establish theoretical analysis on self-supervised learning. [15] establishes the bound between supervised learning and unsupervised learning. [16] understands contrastive learning from a mutual information perspective. Analyzing self-supervised learning from a graph perspective has been proved to be feasible in contrastive learning. [9] analyzes contrastive learning with the spectral graph theory and gives a bound for the downstream performance related to the spectral properties of the augmentation graph. [18] constructs an augmentation graph and analyzes the gap between contrastive learning and supervised learning with it. In this paper, we also try to understand the training process of MAE from the graph perspective, as we find that contrastive

learning and MAE share a significant similarity, *i.e.,* the samples in the same class could generate quite similar views with transformations, including data augmentations used in contrastive learning and mask used in MAE. However, the analysis in this paper is not a trivial solution and there exist some significant differences between contrastive learning and MAE. For example, the existing theory can't explain the relationship between the mask ratio and downstream performance, and the graph of MAE is a directed graph instead of an undirected graph as analyzed in contrastive learning.

Understanding MAE Training Process. Despite the impressive success of MAE, the theoretical understanding of it is still underexplored. [2] is the first work to understand MAE through a mathematical viewpoint, which explains the success of MAE from a perspective of an integral kernel. They analyze the optimizing process of MAE and show the importance of patchifying and the decoder. However, some designs and properties of MAE are still not fully understood. For example, why MAE can learn meaningful representation for downstream tasks with an instance-level reconstruction task and why the mask ratio needs to be set to 0.75? In this paper, we try to further explain this powerful self-supervised training framework from a different perspective, *i.e.,* graph perspective, and answer these questions.

3 Masked AutoEncoding as Implicit Contrastive Learning

3.1 Problem Formulation

We begin by introducing the basic notations and common process of MAE, which consists of two stages, *i.e.,* MAE pretraining and supervised finetuning. In the first stage, we pretrain an encoder-decoder architecture to reconstruct the masked images. In the second stage, we evaluate the quality of the learned representations with labeled dataset $\mathcal{D}_l = \{(x_i, y_i)|x_i \in \mathbb{R}^d, 0 \leq y \leq k\}$ with a classification task.

3.1.1 MAE Pretraining

Assuming access to an unlabeled dataset of natural images $\mathcal{D}_u = \{\bar{x}_i|\bar{x}_i \in \mathbb{R}^d\}$, we denote the natural data distribution as $P_d(\bar{x})$. Masked Autoencoders (MAE) [11] learn data representations with the following procedures.

Pretext Generation - Random Patch-wise Masking. First, for a random image $\bar{x}\in\mathcal{D}_u$, it could be (equivalently) patchified as a matrix, denoted as $\hat{\bar{x}} \in \mathbb{R}^{n\times p}$, where n denotes the number of patches, and p denotes the patch size, *e.g.,* 4×4. Second, according to a certain masking ratio $\rho \in [0, 1]$, we randomly draw a binary mask $m \in \{0, 1\}^n$ for each patch. Denote the number of non-zero elements of m as $|m|$. and selecting patches according to the mask m leads to two complementary views,

$$x_1 = \hat{\bar{x}}[m] \in \mathbb{R}^{|m|\times p}, \quad x_2 = \hat{\bar{x}}[1-m] \in \mathbb{R}^{(n-|m|)\times p}, \tag{1}$$

where $[\cdot]$ denote the index selection function. Overall, we denote this random masking process for generating x_1, x_2 from x as $M(x_1, x_2|\bar{x})$, and their marginal distribution as $M_1(x_1|\bar{x})$ and $M_2(x_2|\bar{x})$, respectively. Here $\mathcal{M}, \mathcal{M}_1, \mathcal{M}_2$ refer to the probability density functions of the corresponding distributions. As long as $\rho \neq 0.5$, the two distributions are different, *i.e.,* $\exists\ x$, *s.t.* $M_1(x|\bar{x}) \neq M_2(x|\bar{x})$. And we denote $p_x \in \mathbb{R}^{n_x\times s}$ as the position information of the view $x \in \mathbb{R}^{n_x\times p}$ with n_x patches, typically represented as Position Encodings (PE).

Reconstruction-based Pretraining. According to He *et al.* [11], the MAE framework is as follows: an encoder f first maps (x_1, p_{x_1}) to a latent feature $z = f(x_1, p_{x_1}) \in \mathbb{R}^{|m|\times q}$. Afterwards, a decoder g takes z_1 and p_{x_2} as the input, and maps to the sample space to get an estimation of x_2, *i.e.,* $\hat{x}_2 = g(z_1, p_{x_2}) = g(f(x_1, p_{x_1}), p_{x_2}) \in \mathbb{R}^{(n-|m|)\times p}$. At last, the encoder-decoder architecture of MAE is learned with a simple ℓ_2 reconstruction loss,

$$\min_{f,g} L_{rec}(\mathcal{D}_u; f, g); \quad L_{rec}(\mathcal{D}_u; f, g) = \mathbb{E}_{P_d(x)}\mathbb{E}_{M(x_1,x_2|x)} \left\|g(f(x_1, p_{x_1}), p_{x_2}) - x_2\right\|^2. \tag{2}$$

3.1.2 Linear Evaluation

After pretraining, we can evaluate the quality of the extracted data representations via a linear classification task. Specifically, assuming access to the labeled dataset $\mathcal{D}_l = \{(x_i, y_i)|x_i \in \mathbb{R}^d, y \in$

$[k]\}$, we train a linear classifier $l : \mathbb{R}^q \to \mathbb{R}^k$ with the encoder held fixed

$$\min_{l} \; L_{cls}(\mathcal{D}_l; f, l); \quad L_{cls}(\mathcal{D}_l; f, l) = \mathbb{E}_{P_d(x,y)} \ell_{ce}(l(f(x)), y), \tag{3}$$

where ℓ_{ce} denotes the cross entropy loss.

3.2 MAE Loss is upper bounded by an Implicit Alignment Loss

Reconstruction Loss is upper bounded by an Implicit Alignment Loss. We begin with an analysis on the reconstruction loss used in the training process of MAE. From its form, we can find that it is a common autoencoding loss that tries to reconstruct samples with an encoder-decoder architecture. As MAE uses the transformer network as the encoder and the decoder, so we can obtain some properties to simplify our subsequent analysis on the training process of MAE. [20] analyzes the universal approximation of the transformer and proposes the following results,

Lemma 3.1 (Theorem 3 [20]). *We denote F_{CD} as the set of all continuous functions that map a compact domain. Let $1 \leq p < \infty$ and $\varepsilon > 0$, then for any given $f \in F_{CD}$, there exists a Transformer network $g \in T_P^{2,1,4}$ such that we have $d_p(f, g) \leq \varepsilon$.*

Combined with the universal approximation property of the transformer network and the encoder-decoder architecture, we propose the following assumption which states there exists a pseudo-inverse encoder of the decoder, which maps from the complementary view (x_2, p_{x_2}) back to the representation space of the encoder.

Assumption 3.2. *For the decoder function $g : \mathbb{R}^{|m| \times q} \times \mathbb{R}^{(n-|m|) \times s} \to \mathbb{R}^{(n-|m|) \times p}$, we denote the pseudo-inverse encoder of it as $f_g : \mathbb{R}^{(n-|m|) \times p} \times \mathbb{R}^{(n-|m|) \times s} \to \mathbb{R}^{|m| \times q}$, satisfying*

$$\forall \, (x, p_x) \in \mathcal{X} \times \mathcal{P}_{\mathcal{X}}, \quad \|g(f_g(x, p_x), p_x) - x\|^2 \leq \varepsilon. \tag{4}$$

Assumption 3.2 characterizes the reconstruction ability of an encoder-decoder architecture. As it does not involve the mask technique, it amounts to a simple autoencoding model, whose reconstruction error can be bounded theoretically and empirically with powerful neural networks [20, 13]. Thus, this assumption is likely to hold. However, we notice that a trivial autoencoding loss can not learn meaningful representations for downstream performance. To verify that MAE reconstruction loss is not only a simple autoencoding loss, we set the mask ratio to 0 and find that the linear accuracy of learned representation on ImageNet-100 decreases to 13.8%. Then we further explore the properties of MAE loss with the above practical assumption,

$$\begin{aligned} L_{rec}(f, g) &= \mathbb{E}_{x_1, x_2} \|g(f(x_1, p_{x_1}), p_{x_2}) - x_2\|^2 \\ &= \mathbb{E}_{x_1, x_2} \|g(f(x_1, p_{x_1}), p_{x_2}) - g(f_g(x_2, p_{x_2}), p_{x_2}) + g(f_g(x_2, p_{x_2}), p_{x_2}) - x_2\|^2 \\ &\leq 2\mathbb{E}_{x_1, x_2} \|g(f(x_1, p_{x_1}), p_{x_2}) - g(f_g(x_2, p_{x_2}), p_{x_2})\|^2 + 2\varepsilon. \end{aligned} \tag{5}$$

From Equation (5), we find the reconstruction loss used in MAE is upper bounded by an implicit alignment loss which pulls in the distance of different masked views encoded by the encoder-decoder architecture. Note that MAE only uses the encoder for downstream tasks, so then we consider the smoothness of the decoder to analyze whether reconstruction loss aligns the different views in the encoder feature space. [12] analyzes the Lipschitz continuity of a transformer network and prove that a transformer network is Lipschitz continuous when we give constraints to the architecture. We also empirically analyze the Lipschitz constant in the input domain of the decoder during the training process of MAE, *i.e.,* $\min_{x_1, x_2 \in \mathcal{D}_u} \frac{\|g(f(x_1, p_{x_1}), p_{x_2}) - g(f_g(x_2, p_{x_2}), p_{x_2})\|^2}{\|f(x_1, p_{x_1}) - f_g(x_2, p_{x_2})\|^2}$ and we find it keepes smaller than 1.42. More details about the empirical verification can be found in Appendix B.1. Combined with the empirical verification, it's practical to assume that the decoder g is L-Lipschitz, *i.e.,*

Assumption 3.3. *For the decoder function g, we assume there exists a constant L, satisfying*

$$\forall \, (x_1, x_2), \quad \|g(x_1) - g(x_2)\|^2 \leq L\|x_1 - x_2\|^2. \tag{6}$$

Theorem 3.4. *Under Assumptions 3.2 & 3.3, the MAE reconstruction loss can be upper bounded by the alignment loss between $\tilde{x}_1$ and $\tilde{x}_2$, where $\tilde{x} = [x, p_x]$ denotes the position-augmented input:*

$$L_{rec}(f, g) \leq -2L \cdot \mathbb{E}_{\tilde{x}_1, \tilde{x}_2}(f(\tilde{x}_1)^\top f_g(\tilde{x}_2)) + 2\varepsilon + C. \tag{7}$$

Here ε is the approximation error of f_g and C is a constant.

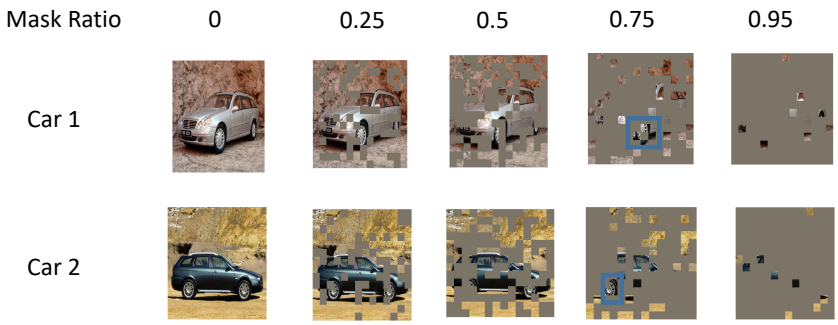


Figure 2: Appropriate mask ratio can generate similar views from different samples in the same class.

Intuitively, Theorem 3.4 shows that the reconstruction loss of MAE implicitly minimizes the distance of different masked views of the same sample in the encoder features space. It can be seen that this is similar to the alignment objective that appears in contrastive (as well as non-contrastive) methods, *i.e.,* $L(f,g) = -2\mathbb{E}_{x_1,x_2} f(x_1)^\top g(x_2)$. However, there exist obvious differences as the views used in MAE is asymmetric with different mask ratio while the views in canonical contrastive learning are symmetric with data augmentations. Meanwhile, MAE could avoid latent collapse because f_g is always a nontrivial target.

Suggested Uniformity-promoting MAE Loss. With the proof above, we can find that the MAE loss is upper bounded by an implicit alignment loss, which aligns features of the masked and unmasked views. [17] understands the alignment loss from two properties, *i.e.,* alignment and uniformity. As proposed in [17], without keeping the uniformity of the features, the cluster performance of the same class will be significantly hurt. So we add a regularizer term to ensure the uniformity of the encoder to improve the performance of MAE and propose our suggested Uniformity-promoting MAE (U-MAE) loss,

$$L_{\text{U-MAE}}(f,g) = \mathbb{E}_{\tilde{x}_1,\tilde{x}_2}\|g(f(x_1,p_{x_1}),p_{x_2}) - x_2\|^2 - \lambda \cdot \mathbb{E}_{\tilde{x}_1,\tilde{x}_2^-}\|f(\tilde{x}_1) - f_g(\tilde{x}_2^-)\|^2, \tag{8}$$

where $\tilde{x}_1$, $\tilde{x}_2$ is the unmasked and masked views of the same sample, *i.e.,* the positive samples, and $\tilde{x}_2^-$ is the masked view of an independently drawn sample, *i.e.,* the negative sample. As the implicit reverse encoder f_g is a hypothetical function and hard to obtain, we replace it with the encoder f, and obtain the following practical U-MAE loss with symmetric uniformity regularizer:

$$L^{\text{sym}}_{\text{U-MAE}}(f,g) = \mathbb{E}_{\tilde{x}_1,\tilde{x}_2}\|g(f(x_1,p_{x_1}),p_{x_2}) - x_2\|^2 - \lambda \cdot \mathbb{E}_{\tilde{x}_1,\tilde{x}_1^-}\|f(\tilde{x}_1) - f(\tilde{x}_1^-)\|^2, \tag{9}$$

where $\tilde{x}_1^-$ is an independently drawn unmasked view. Intuitively, a large distance between encoder outputs can also effectively promote feature diversity.

4 Generalization Theory

4.1 Spectral Formulation

Based on our above analysis of MAE loss, we know that MAE loss is upper bounded by an implicit alignment loss. In this section, we revisit the training process of MAE from a spectral graph perspective and find that MAE reconstruction loss is closely related to an asymmetric matrix factorization problem. Built upon this connection, we establish theoretical guarantees on the downstream performance and characterize its influencing factors.

Constructing the Directed Augmentation Graph. Following [9], we define the population augmentation graph $G(\mathcal{X}, A)$, where $\mathcal{X}$ denotes all masked views of size $N = |\mathcal{X}|$, and $A = (w_{xx'}) \in \mathbb{R}^{N\times N}$ denotes the adjacency matrix filling with edge weight $w_{xx'}$ for two masked views $x, x' \in \mathcal{X}$. Specifically, we define the weight $w_{xx'}$ as the marginal probability of generating the pair x and x' from the random natural data

$$w_{xx'} = \mathbb{E}_{\bar{x}\sim\mathcal{P}_d}[M_1(x|\bar{x})M_2(x'|\bar{x})]. \tag{10}$$

Intuitively, the population augmentation graph formalizes the process of how mask technique bridges different samples in the same ground-truth class. For example, the tires of two different cars could be quite similar and can be seen as generated from the same sample, then some masked views of these two cars will have large weights between them. Obviously, the samples of the same class will have larger weights then the samples of different classes. Note that different from the undirected augmentation graph considered in [9], the graph $G(\mathcal{X}, w)$ is directed, *i.e.,*

$$\mathbb{E}_{\bar{x}\sim\mathcal{P}_d}[A_1(x|\bar{x})A_2(x'|\bar{x})] = w_{xx'} \neq w_{x'x} = \mathbb{E}_{\bar{x}\sim\mathcal{P}_d}[A_1(x'|\bar{x})A_2(x|\bar{x})]. \tag{11}$$

Accordingly, we can normalize the directed adjacency matrix A as $\bar{A} := D_{row}^{-1/2} A D_{col}^{-1/2}$, where D_{row}, D_{col} are diagonal matrices with $\forall x \in \mathcal{X}$, $(D_{row})_{xx} = w_{x\cdot} := \sum_{x'} w_{xx'}$, $(D_{col})_{xx} = w_{\cdot x} := \sum_{x'} w_{x'x}$. Because $\bar{A}$ is asymmetric, generally $w_{x\cdot} \neq w_{\cdot x}$.

Asymmetric Matrix Factorization. With the directed augmentation graph, we revisit our U-MAE loss and find it can be upper bounded by an Asymmetric Matrix Factorization loss.

Theorem 4.1. *For the augmentation graph adjacency matrix A, we consider two matrices $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{k\times n}$, and a decomposition objective $L_{mf}(U, V) = \|\bar{A} - UV^\top\|_F^2$. We have*

$$L_{U-MAE}(f, g) \leq 2L_{mf}(f, f_g) + 2\varepsilon + C, \tag{12}$$

where ε is the approximation error of the transformer network.

Built upon this connection, in the next step, we will discuss the downstream performance by analyzing the asymmetric matrix decomposition objective. To achieve this, we first introduce two common assumptions adopted in previous work [9, 18], one on the label consistency of masking and one on the expressivity of the hypothesis class. Both are likely to hold in practice as long as we choose proper mask ratio and architecture for MAE.

Assumption 4.2 (Labels are recoverable from masked views)**.** *Let $\bar{x} \in \mathcal{D}_u$ be a sample and $y(\bar{x})$ be its label. For a masked view x_1 of x, we assume there exists a classifier f that can predict $y(\bar{x})$ given x with error at most α. That is, $f(x_1) = y(\bar{x})$ with probability at least $1 - \alpha$.*

Assumption 4.3 (Expressivity of the hypothesis class)**.** *Let $\mathcal{F}, \mathcal{G}$ be the hypothesis classes of the encoder f and the decoder g, respectively. We assume that at least one of the global minima of $L_{\text{U-MAE}}(f, g)$ belongs to $\mathcal{F}$ and $\mathcal{G}$.*

In the following, combined with our analysis on Theorem 4.1, we establish a bound between the downstream performance of MAE and the spectral property of the augmentation graph. The details of the proof can be found in Appendix A.3.

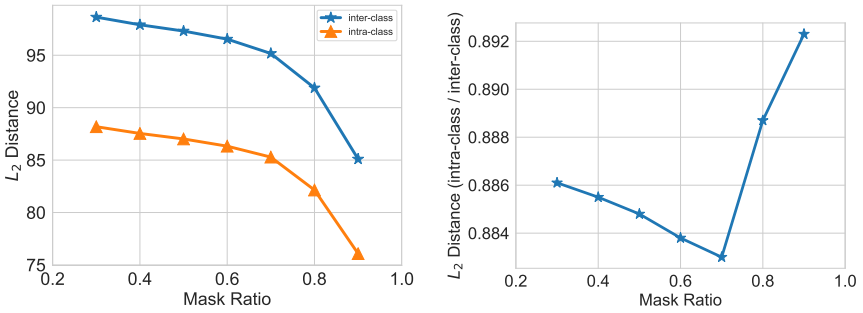
Theorem 4.4. *Under Assumptions 3.2, 3.3, 4.2 & 4.3, let $\mathcal{L}$ be the Laplacian matrix of the adjacent matrix A and $\lambda_1, \lambda_2, \cdots, \lambda_k, \lambda_{k+1}$ be be the $k+1$ smallest eigenvalues of $\mathcal{L}$. $\zeta(f^\star_{\text{U-MAE}})$ is denoted as the downstream error for the minimizer of MAE loss $f^\star_{\text{U-MAE}}$, we have*

$$\zeta(f^\star_{\text{U-MAE}}) \leq O(L\sqrt{2\alpha}/\lambda_{k+1}) + 2\varepsilon. \tag{13}$$

From Theorem 4.4, we can conclude that the downstream performance can be minimized with a small Lipschitz constant L, a small decoder inversion error ε, a small the labeling error α, and a large $k+1$-th eigenvalue of the augmentation graph λ_{k+1}. In other words, MAE with better landscape smoothness and invertible decoder will be likely to have a smaller downstream error, and in the meantime the masking ratio ρ should be properly chosen such that the labeling error α is small and λ_{k+1} is large. The last condition seems less intuitive. In fact, according to spectral graph theory [4], a large eigenvalue λ_{k+1} stems from a strong connectivity of the augmentation graph $\mathcal{G}$. Thus, MAE often requires a large masking ratio to bridge different samples together. Meanwhile, as illustrated in Figure 2, a too large mask ratio, *e.g.,* $\rho = 0.95$, could make it hard to recover the label from the masked views, which results in a large α. As a result, Theorem 4.4 suggests that there exists a tradeoff in the mask ratio ρ, and we should choose a properly large one, *e.g.,* $\rho = 0.75$.

4.2 Empirical Investigation of the Directed Augmentation Graph of MAE

Properties. As discussed in [18], contrastive learning bridges the samples in the same class by the similar views generated by appropriate data augmentations, comparing to the augmentation graph



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

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

Figure 3: The influence of mask ratio on (a) the average $L2$ distance on ImageNet-100 between intra-class samples and inter-class class samples, and (b) the relative distance between intra and inter-class samples.

used in MAE, we can find that both of them align the intra-class samples by transforming different anchor samples into similar views. However, although they have a lot in common, there exist obvious differences between contrastive learning and MAE. Firstly, as the mask ratio is not equal to 0.5, the two views of an anchor sample in MAE are asymmetric, which means the augmentation graph changes from an undirected graph in contrastive learning into a directed graph in MAE. So we can not directly use the theoretical framework designed for analyzing the undirected augmentation graph in contrastive learning and we find it can be analyzed with an Asymmetric Matrix Factorization objective. And as the mask ratio of similar views of different anchors should be equal, the augmentation graph of MAE becomes a bipartite graph, which means views of different samples with the same mask ratio can't be connected directly and they need a middleman between them. What's more, the two views of the same anchor in contrastive learning is independent while the two views in MAE loss are dependent and contain the whole anchor, which means the two views in MAE training can represent most of the features of the sample.

The Necessity of High Mask Ratio. There exists a difficulty in understanding why the training process of MAE needs a high mask ratio. We find that a major reason is that the higher mask ratio will generate more similar views of intra-class samples. For example, in Figure 2, we can find that when the mask ratio increases to 75%, the views of two different cars can be concentrated on the features of similar tires. To verify our perspective, we conduct an experiment on images of CIFAR-10 with different masked ratios. We evaluate the distance between two images by calculating the l_2 distance between the patches of them. As shown in Figure 3(a), we can find that the average distance of intra-class images decreases with the enhancement of mask ratio, which means a higher mask ratio leads to generating more edges between intra-class samples.

Drawbacks of Too High Mask Ratio. On the other hand, as the highest mask ratio can not lead to the highest downstream performance of MAE, we find it is related to the decreased size of the support set of an anchor, *i.e.,* some transformed views of an anchor will be very similar. Indeed, as shown in Figures 3(a), the average L_2 distance between inter-class samples and overall samples decreases as the mask ratio increases, which means when we take an excessive mask ratio, numerous views will be merged and the support set of the anchors will become smaller. And the views with an extremely high mask ratio will lose the feature of their class. For example, as shown in Figure 2, we can find that when the mask ratio arrives at 95%, the features of the car disappear.

The Sweet Spot for Mask Ratio. The discussion above reveals that either too small or too high mask ratios can result in bad features, and a good mask ratio should be selected such that intra-class distance is relatively high while inter-class distance is relatively low. Motivated by this, we plot the relative distance (intra-class over inter-class) in Figure 3(b), and we can find that the relative distance decreases first with larger 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

Table 1: Linear evaluation accuracy and finetune accuracy on CIFAR-10 and ImageNet-100 pretrained with MAE loss and U-MAE loss with different ViT backbones. The uniformity regularizer term significantly improves the linear evaluation performance of MAE loss with different backbones across different real-world datasets without hurting the performance of finetune accuracy.

	Objective	CIFAR-10		ImageNet-100	
BackBone		ViT-Tiny	ViT-Base	ViT-Base	ViT-Large
Linear Evaluation	MAE	52.0	59.9	37.5	39.5
	U-MAE	69.4	72.0	56.3	61.4
Finetune Results	MAE	89.6	90.7	86.9	87.3
	U-MAE	89.4	90.8	86.8	87.3

large mask ratio. The sweet spot lies in $\rho = 0.7$, which is pretty close to the optimal masking rate of MAE is $\rho = 0.75$ [11]. This shows that our theoretical analysis of the effect of MAE ratio agrees surprisingly well with the practice of MAE.

5 Experiments

5.1 Evaluation on Benchmark Datasets

We conduct experiments of our U-MAE loss on various real-world datasets, including CIFAR-10 [14] and ImageNet-100 [5]. In the following, we introduce our empirical setup on CIFAR-10 and ImageNet-100.

Setup. As our proposed loss is a promoting MAE loss, we mainly follow the basic setup of MAE. We use ViT as the encoder and use the flexible decoder as proposed in MAE. To further present the performance of our loss, we carry out our experiment on different variants of ViT, *i.e.,* ViT-Tiny, ViT-Base and ViT-Large. We follow the default setting as proposed in [11]. We use the recommended mask ratio, 75%. For the uniformity term of our proposed loss, we set the coefficient of the uniformity term to 0.001. For ImageNet-100, we pretrain the model for 200 epochs with batch size 256 and weight decay 0.05. And for CIFAR-10, we pretrain the model for 2000 epochs with batch size 4096 and weight decay 0.05. On the stage of supervised finetuning, we train a supervised classifier on the pretrained encoder. We conduct both linear evaluation and non-linear finetuning on the unsupervised pretrained encoder. For linear evaluation, we train a linear classifier on the frozen pretrained encoder. As for non-linear finetuning, we train both the pretrained encoder and the linear classifier with Cross Entropy loss.

Results. In Table 1, we compare the linear evaluation results between original MAE loss and U-MAE loss on different benchmarks, including CIFAR-10 and ImageNet-100, and different backbones, including ViT-Tiny, ViT-Base and ViT-Large. We find that our promoting loss increases 14.71% for linear evaluation results on CIFAR-10 with two different backbones and increases 20.35 % on ImageNet-100 with two different backbones. And in Table 1, we also present the comparison of finetuning results. We find that our proposed loss will not hurt the performance of finetuning results of MAE. Our results empirically verify the effectiveness of our loss and show that the U-MAE loss achieves better performance than the original MAE loss across different datasets and different backbones.

5.2 Empirical Understanding and Ablation Study.

Visualization of Representations. To intuitively understand the improvement of our U-MAE loss on clustering intra-class samples, we use t-SNE to visualize the representations trained with MAE loss and our proposed loss on ten random class of ImageNet-100 datasets 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 (representing for hens) and the gray class (representing for indigo birds) are separated from others. We note most of other classes are the animals living in the oceans) while these two types are more like the birds (living on the land or the sky). So these two classes are more easier to be distinguished, especially with our uniformity regularizer term.

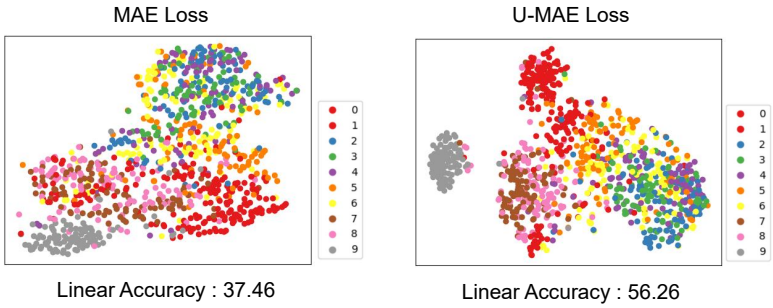
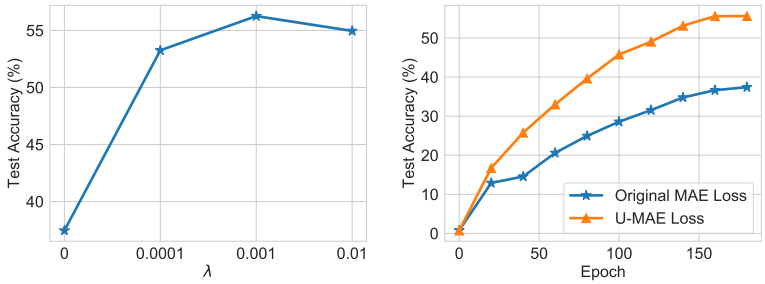


Figure 4: Visualization of representations on random 10 classes of ImageNet (0-9 classes described in Appendix B.4) trained with MAE loss and our U-MAE loss. Our loss significantly improves the class-clustering performance of the encoder.



(a) Different coefficient of the uniformity regularizer term.

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

Figure 5: (a) The linear evaluation results trained with different coefficients of the uniformity regularizer term of our U-MAE loss. It shows that an appropriate coefficient is important for our proposed loss. (b) The linear evaluation results during the training process. It shows that our proposed loss improve the downstream performance of MAE with different training time.

Different Coefficients of the Regularizer Term. The most important hyper-parameter of our proposed loss is the coefficient of the uniformity regularizer term. In Figure 5, 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 find that when the coefficient is 0, our proposed loss is an original MAE loss. The downstream performance increases when the coefficient increases from 0 to 0.001. However, the overlarge coefficient will also hurt the performance of our proposed loss as the task of MAE will be overlooked.

Training process. To further compare the performance between the original MAE loss and our U-MAE loss, we report the linear evaluation accuracy on ImageNet-100 during the training process. In Figure 5, we show the linear evaluation results during the training process. We can observe that our proposed loss improves the performance of MAE with all different training epochs, which verifies that our proposed loss is a promoting loss of MAE for both short and long time learning.

6 Conclusion

In this paper, we propose a new theoretical understanding of MAE. With the analysis of MAE loss, we find that MAE loss is upper bounded by an implicit alignment loss. Then we explain the training process of MAE from a graph perspective and establish a guarantee for its downstream performance. Based on our theory, we propose an Uniformity-promoting MAE (U-MAE) loss and verify that it can significantly improve the downstream performance of MAE across different benchmark datasets, including CIFAR-10 and ImageNet-100.

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

[4] Fan Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

[5] 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.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[8] 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.

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

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

[11] 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.

[12] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *ICML*, 2021.

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

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

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

[16] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020.

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

[18] 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.

[19] 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*, pages 9653–9663, 2022.

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

[21] 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)? [No] Release upon acceptance.
   (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]