i-MAE: Are Latent Representations in Masked Autoencoders Linearly Separable?

Kevin Zhang[†,‡,*], Zhiqiang Shen[§,*]
[†]Peking University [‡]KNQ.AI [§]Mohamed bin Zayed University of AI

Abstract

Masked image modeling (MIM) has been recognized as a strong self-supervised pre-training approach in the vision domain. However, the mechanism and properties of the learned representations by such a scheme, as well as how to further enhance the representations are so far not well-explored. In this paper, we aim to explore an interactive Masked Autoencoders (i-MAE) framework to enhance the representation capability from two aspects: **(1)** *employing a two-way image reconstruction and a latent feature reconstruction with distillation loss to learn better features;* **(2)** *proposing a semantics-enhanced sampling strategy to boost the learned semantics in MAE. Upon the proposed i-MAE architecture, we can address two critical questions to explore the behaviors of the learned representations in MAE:* **(1)** *Whether the separability of latent representations in Masked Autoencoders is helpful for model performance? We study it by forcing the input as a mixture of two images instead of one.* **(2)** *Whether we can enhance the representations in the latent feature space by controlling the degree of semantics during sampling on Masked Autoencoders? To this end, we propose a sampling strategy within a mini-batch based on the semantics of training samples to examine this aspect. Extensive experiments are conducted on CIFAR-10/100, Tiny-ImageNet and ImageNet-1K datasets to verify the observations we discovered. Furthermore, in addition to qualitatively analyzing the characteristics of the latent representations, we examine the existence of linear separability and the degree of semantics in the latent space by proposing two evaluation schemes. The surprising and consistent results across the qualitative and quantitative experiments demonstrate that i-MAE is a superior framework design for understanding MAE frameworks, as well as achieving better representational ability.*

1. Introduction

Self-supervised learning aims to learn representations from abundant unlabeled data for benefiting various downstream tasks. Recently, many self-supervised approaches have been proposed in the vision domain, such as pre-text based methods [8, 11, 32], contrastive learning with siamese networks [4, 13, 15, 23], masked image modeling (MIM) [2, 14, 28, 29], etc. Among them, MIM has shown a preponderant advantage in finetuning performance, and the representative method Masked Autoencoders (MAE) [14] has attracted much attention in the field. A natural question is then raised: *Where are the benefits of the finetuning transferability to downstream tasks from in MAE-based training?* This motivates us to develop a framework to shed light on the reasons of MAE's superior latent representations. Further, we can explore the feasibility of leveraging the discoveries to continue enhancing the representation ability of MAE. As the understanding of MAE framework remains under-studied, it is crucial to explore it more in a specific and exhaustive way for better performance.

Intuitively, a good representation should be separable and contain enough semantics from its input, so that it can have a qualified ability to distinguish different classes with

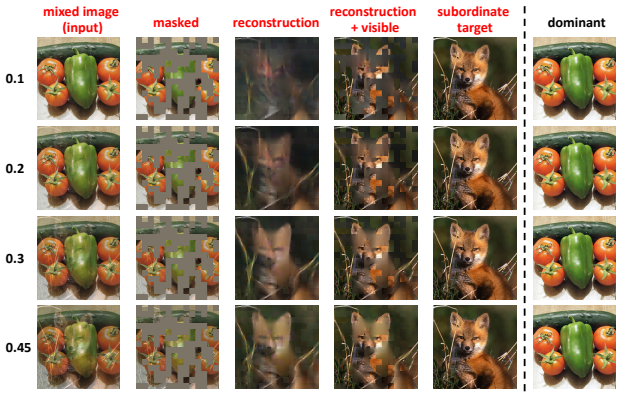

Figure 1. Reconstruction results from the subordinate branch of *i-MAE* on ImageNet-1K validation images with different mixing coefficients α (listed on the left). i-MAE is pre-trained with linearly mixed input reconstruction loss on both inputs. Visually, i-MAE predictions reflect features of $\mathbf{I}_s$ even at low mixture coefficients and still reconstruct the subordinate image well, whereas at 0.45 which is more challenging, the reconstructions show elements like colors from the dominant image but the content still matches the target. More visualizations are provided in Appendix.

*The two authors have equal contribution to this work. Zhiqiang Shen is the corresponding author. Code is available on GitHub.

better performance on downstream tasks. This inspires us to question "can we utilize the idea of separability and more semantics to enhance the representation capability for Masked Autoencoders?" This is in general difficult as how to evaluate the separability and the degree of semantics on the latent features is not clear thus far. Moreover, the success of an Autoencoder *compressing* the information from input by reconstructing itself has been well-recognized only in practice; the explanation and interpretability of the features learned from such approaches is still under-explored.

To address the difficulties of identifying separability and semantics in latent features, we first propose a novel framework, i-MAE, upon vanilla MAE. It consists of a mixture-based masked autoencoder branch for disentangling the mixed representations by linearly separating two different instances, and a pre-trained vanilla MAE as guidance for distilling the disentangled representations. An illustration of the overview framework architecture is shown in Fig. 2. This framework is designed to answer two interesting questions: **(1)** Are the latent representations in Masked Autoencoders *linearly separable*? More importantly, how can we utilize this property to learn stronger and better features? **(2)** Whether we can enhance the representations for Masked Autoencoders by leveraging *more semantics* through a sampling strategy within a mini-batch? We attribute the superior representation capability of MAE to its learned separable features for downstream tasks with enough semantics.

In addition to qualitative studies, we develop two quantitative evaluation schemes to address the two questions quantitatively. In the first evaluation, we employ several distance measurements of root ℓ_2, *R-Squared* value after regression and *cosine*-similarity from high-dimensional Euclidean spaces to measure the similarity between i-MAE's disentangled feature and the "ground-truth" feature from pre-trained MAE on the same image. In the second evaluation, we control different ratios of semantic classes as a mixture within a mini-batch and evaluate the finetuning and linear probing results of the model to reflect the learned semantic information. More details are in Sec. 3.2 and 3.3.

We conduct extensive experiments on different scales of datasets: CIFAR-10/100, Tiny-ImageNet and ImageNet-1K to demonstrate the effectiveness of i-MAE with better capability, studying the linear separability and the degree of semantics in the latent representations. We further provide both qualitative and quantitative results to explain our observations and discoveries. The characteristics we observed in latent representations according to our proposed i-MAE framework are: **(1)** i-MAE learned feature representation has great linear separability for its input data, which can be beneficial for linear probing and finetuning tasks. **(2)** Though the training scheme of MAE is different from instance classification pre-text in contrastive learning, its representation still encodes sufficient semantic information from input data. Moreover, *mixing the same-class images as the input training samples substantially improves the quality of learned features*. **(3)** We can reconstruct the individual images from a mixture with i-MAE effortlessly, even if it is the subordinate part. To the best of our knowledge, this is the pioneering study to explicitly explore the separability and semantics of a mixed MAE scheme with extensive well-designed qualitative and quantitative experiments.

Our contributions in this work are three-fold:

- We propose an *i-MAE* framework with two-way image reconstruction and latent feature reconstruction with a distillation loss to boost the representation capability inside the MAE framework and explore the understanding of mechanisms and properties of learned representations.
- We evaluate our dual-reconstruction framework and find better linear separability of features which are more interpretable. We further introduce a semantics-enhancement sampling strategy, a straightforward yet effective scheme to increase the quality of learned features.
- We conduct extensive experiments on various scales of datasets: CIFAR-10/100, Tiny-ImageNet and ImageNet-1K, and we provide sufficient qualitative and quantitative results to verify the effectiveness of proposed framework.

2. Related Work

Masked image modeling. Motivated by masked language modeling's success in language tasks [7, 24], Masked Image Modeling (MIM) in vision learns representations from images corrupted by masking. State-of-the-art results on downstream tasks are achieved by several approaches. BEiT [2] proposes to recover discrete visual tokens and SimMIM [29] performs pixel-level reconstruction. Recently, MAE [14], which recovers pixels from a high masking ratio, has been shown capable of learning robust representations [1, 10, 33]. In this work, we focus on designing a mixture strategy to enhance representations learned by MAE [14].

Image mixtures. Widely adopted mixture methods in visual supervised learning include Mixup [31] and Cutmix [30]. However, these methods require ground-truth labels for calculating mixed labels; in this work, we adapt Mixup to our unsupervised framework by formulating losses on dual reconstructions. On the other hand, in recent visual self-supervised learning literature, joint embedding methods and contrastive learning approaches such as MoCo [13], SimCLR [4], and more recently UnMix [25] have acquired success and predominance in mixing visual inputs, promoting instance discrimination by aligning features of augmented views of the same image. Related to our work, [20] have proposed to use cutmix in MIM, replacing mask tokens with visible tokens of another image and performing dual reconstructions. Contrarily, we conduct image mixtures at the pixel level rather than token-level, with the

unique advantage of being more interpretable with latent decomposition.

Invariance and disentanglement of representation learning in Autoencoders. Representation learning focuses on the properties of features learned by the layers of deep models while remaining agnostic to the particular optimization process. Invariance and disentanglement are two commonly discussed factors that occur in data distribution for representation learning [21, 22]. Autoencoders are classical generative unsupervised representation learning frameworks based on image reconstruction as loss function, learning both the mapping of inputs to latent features and the reconstruction of the original input. In this work, we focus on the latent disentanglement where one feature is correlated or connected to other vectors in the latent space in Masked Autoencoder. The motivation for learning disentangled features in Autoencoders is for achieving interpretability [3, 5] and for intuitive explanations.

3. i-MAE

In this section, we first introduce an overview of our proposed framework. Then, we present our components in detail, including our mixture input, two-branched reconstruction via a shared-decoder, and our patchwise-distillation module. Ensuingly, we elaborate on the metric we propose to evaluate the improved linear separability in i-MAE and the sampling strategy to enhance the degree of semantics in our mixtures, as well as broadly discussing the observations and discoveries.

3.1. Framework Overview

As shown in Fig 2, our framework consists of three submodules: **(1)** a mixture encoder module that takes the masked mixture image as the input and output mixed features; **(2)** a disentanglement module that splits the mixed feature into the individual ones; **(3)** a MAE teacher module that provides the pre-trained embedding for guiding the splitting process in the disentanglement module.

3.1.1 Components

Input Mixture with MAE Encoder. Inspired by Mixup [31], we use an unsupervised mixture of inputs formulated by $\alpha * \mathbf{I}_1$ and $(1-\alpha) * \mathbf{I}_2$, where $\mathbf{I}_1, \mathbf{I}_2$ are the input images. Essentially, our encoder extrapolates mixed features from a tiny fraction (e.g., 25%) of visible patches, then we linearly project it to represent both input images separately. Formally, the mixed image is:

$$\mathbf{I}_m = \alpha * \mathbf{I}_1 + (1-\alpha) * \mathbf{I}_2, \tag{1}$$

where α is the coefficient to mix two images following a Beta distribution. The encoder of i-MAE directly follows MAE [14], generating input tokens from images. The

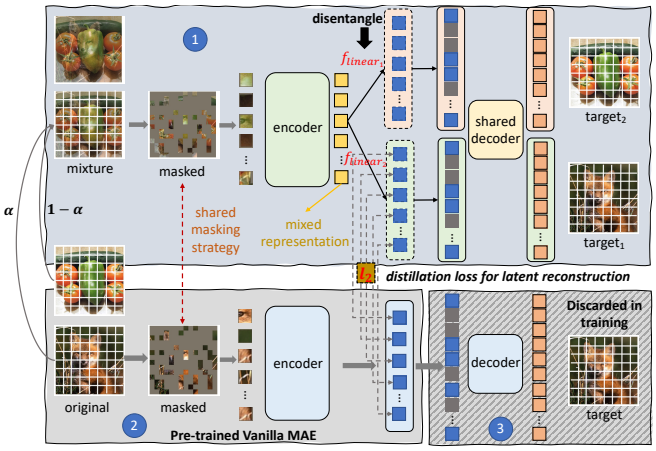

Figure 2. Framework overview of our *i-MAE*. ① is the main branch that consists of a mixture encoder, a disentanglement module, and a two-way image reconstruction module. ② is the encoder part of a pre-trained vanilla MAE for distillation purposes (i.e., latent reconstruction). ③ is the decoder part in MAE and is discarded in training.

mixed input is divided into non-overlapping patches and then goes through the embedding procedure. The same masking strategy is used in both the teacher (Vanilla MAE) and student (i-MAE) models.

Two-branch Masked Autoencoders with Shared Decoder. Although sufficient semantic information from both images is embedded in the mixed representation to reconstruct both images, the vanilla MAE cannot by itself associate the entangled features with either input. The MAE structure does not retain identification information (e.g., order or positional information) about the two inputs that are mixed in image space, i.e., the model cannot tell which of the two images to reconstruct to, since both are sampled from the same distribution and mixed randomly. The consequence is that both reconstructions look identical to each other and fail to look similar to either original input.

Similar to how positional embeddings are needed to explicitly encode spatial information, i-MAE implicitly encodes the semantic difference between the two inputs by using a dominant and subordinate mixture strategy. In practice, we train the linear separation layers to distinguish between the dominant input $\mathbf{I}_d$ (higher mix factor) and the subordinate input $\mathbf{I}_s$ (lower mix factor).

Two-way Image Reconstruction Loss. Formally, we build our reconstruction loss to recover individual images from a mixed input, which is first fed into the encoder to generate mixed features:

$$\mathbf{h}_m = \mathbf{E}_{\text{i-MAE}}(\mathbf{I}_m), \tag{2}$$

where $E_{\text{i-MAE}}$ is i-MAE's encoder, $\mathbf{h}_m$ is the latent mixed representation. Then, we employ two non-shareable linear embedding layers to separate the mixed representation into

individual ones:

$$\begin{aligned} \mathbf{h}_1 &= \mathcal{F}_1(\mathbf{h}_m), \\ \mathbf{h}_2 &= \mathcal{F}_2(\mathbf{h}_m), \end{aligned} \tag{3}$$

where $\mathcal{F}_1$, $\mathcal{F}_2$ are two linear layers with different parameters for disentanglement, and $\mathbf{h}_1$ and $\mathbf{h}_2$ are corresponding representations. After that, we feed the individual representations into the shared decoder with the corresponding reconstruction losses:

$$\begin{aligned} \mathcal{L}_{\text{recon}}^{\mathbf{I}_1} &= \mathbb{E}_{\mathbf{I}_1 \sim p(\mathbf{I}_1)}[\|\mathbf{D}_{\text{shared}}(\mathbf{h}_1) - \mathbf{I}_1\|_2], \\ \mathcal{L}_{\text{recon}}^{\mathbf{I}_2} &= \mathbb{E}_{\mathbf{I}_2 \sim p(\mathbf{I}_2)}[\|\mathbf{D}_{\text{shared}}(\mathbf{h}_2) - \mathbf{I}_2\|_2], \end{aligned} \tag{4}$$

where $p(I_n)$ is the sample distribution of input images.

We have shown that our encoder learns to embed representations of both images. We also examined reconstructing only the subordinate image $\mathbf{I}_s$ to prevent the $\mathbf{I}_d$ from guiding the reconstruction. We empirically verified that a single branch is sufficient for the proposed framework to converge training and reconstruct the subordinate target. We provide comparisons with the representational abilities of the single branch i-MAE in the experiments. By default, we choose the double-branched i-MAE for stronger representation performance. Essentially, successful reconstructions from only the $\mathbf{I}_s$ prove that representations of both images can be learned and that the subordinate image is not filtered out as noise.

Concretely, through an unbalanced mix ratio and a reconstruction loss targeting only one of the inputs, our framework encodes sufficient information for i-MAE to linearly map the input mixture to two outputs.

Patch-wise Distillation Loss for Latent Reconstruction. With the linear separation layers and an in-balanced mixture, the i-MAE encoder is presented with sufficient information about both images to perform visual reconstructions. However, information is inevitably lost during the mixing process, harming the value of the learned features in downstream tasks such as classification. To mitigate such an effect, we propose a knowledge distillation module for not only enhancing the learned features' quality, but also demonstrating that a successful distillation can evidently prove the linear separability of our features.

Intuitively, MAE's features can be regarded as "ground-truth" and i-MAE learns features distilled from the original MAE. Specifically, our loss function computes ℓ_2 loss between disentangled representations and original representations to help our encoder learn useful features of both inputs. Our Patch-wise latent reconstruction loss can be formulated as:

$$\begin{aligned} \mathcal{L}_{\text{recon}}^{\mathbf{h}_1} &= \mathbb{E}_{\mathbf{h}_1 \sim q(\mathbf{h}_1)}[\|\mathbf{E}_{\text{p-MAE}}(\mathbf{I}_1) - \mathbf{h}_1\|_2], \\ \mathcal{L}_{\text{recon}}^{\mathbf{h}_2} &= \mathbb{E}_{\mathbf{h}_2 \sim q(\mathbf{h}_2)}[\|\mathbf{E}_{\text{p-MAE}}(\mathbf{I}_2) - \mathbf{h}_2\|_2], \end{aligned} \tag{5}$$

where $E_{\text{p-MAE}}$ is the pre-trained MAE encoder.

Semantics-enhanced Sampling. Our *semantics-enhanced sampling* is a data sampling strategy that introduces significantly more image mixtures belonging to the same class into the training process, which boosts the semantics learned by the double-branched i-MAE. Specifically, we select training instances from the same classes following different distributions to constitute an input mixture as follows:

$$p = \mathcal{F}_m(\mathbf{I}_{c_a} + \mathbf{I}_{c_b}), \tag{6}$$

where $\mathcal{F}_m$ is the backbone network for mixture input and p is the corresponding prediction. $\mathbf{I}_{c_a}$ and $\mathbf{I}_{c_b}$ are the input samples, and c_a, c_b have a certain percentage r that belongs to the same category. For instance, if $r = 0.1$, it indicates that 10% images in a mini-batch are mixed with the same class. When $r = 1.0$, all training images will be mixed with another one from the same class. We train the model with a fixed r, and find that the intuitive method can effectively enhance learned semantics.

3.2. Linear Separability

In this section, we explain the approaches we utilize to measure linear separability of the features learned in i-MAE. Many classical works on Autoencoders have demonstrated the improved interpretability from disentanglement [3] and especially linear disentanglement [16, 26]. In i-MAE, the motivation behind our pre-training strategy is for more linearly separable features.

For i-MAE to reconstruct both the subordinate and dominant image from a potentially linear mixture, not only should the encoder be general enough to retain information from both inputs, but it also needs to generate embeddings that are specific enough for the decoder to distinguish them into their pixel-level forms. A straightforward interpretation of how i-MAE fulfills both conditions is that the latent mixture $\mathbf{h}_m$ is a linear combination of features that closely relate to $\mathbf{h}_1$ and $\mathbf{h}_2$, e.g., in a linear relationship. To verify this explanation, we employ a linear separability metric to experimentally observe such a behavior.

Metric of Linear Separability. A core contribution of our i-MAE is that the framework learns features that are better linearly disentangled, and we provide tools to evaluate this observation. In general, linear separability is a property of two sets of features that can be separated into their respective sets by a hyperplane. In our example, the set of latent representations $\mathbf{H}_1$ and $\mathbf{H}_2$ are linearly separable if there exists $n+1$ real numbers $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n, \mathbf{b}$, such that every $\mathbf{h} \in \mathbf{H}_1$ satisfies $\sum \mathbf{w}_i \mathbf{h}_i > \mathbf{b}$ and every $\mathbf{h} \in \mathbf{H}_2$ satisfies $\sum \mathbf{w}_i \mathbf{h}_i < \mathbf{b}$. It is a common practice to train a linear classifier (e.g., SVM [17]) or a linear regressor [9, 12] for evaluating if the two sets of features are linearly separable.

To measure the linear relation of disentanglement, we train a linear regressor with with ℓ_1 regularization (lasso

penalty) between the disentangled features of the subordinate image $\mathbf{I}_s$ and the vanilla MAE features of the same input $\mathbf{I}_s$. Intuitively, since the disentangled features without constraints will be far from the features of the vanilla MAE model, we utilize a linear regressor to fit the disentangled features to the vanilla features for comparisons. To quantitatively measure the linear separability of i-MAE, we utilize a variety of scores and distances to evaluate the manifold correlations, including *Normalized Root Mean Square Error (NRMSE)*, *Coefficient of Determination (R-squared R^2)*, and *Cosine Similarity*. Among them, NRMSE and R-squared are used to evaluate the correlation of regression, and cosine distance can measure how close the disentangled features are with the original features when mapped to the same latent space.

3.3. Semantics

Enhanced Semantics. In our mixture strategy and double reconstructions, the representational ability learned is affected by the mixing strategy; hence, we improve our learning framework with semantics-enhanced sampling, an approach that samples image mixtures pertaining to the same class. Intuitively, the disentanglement of features from the same class is more difficult than segmenting different classes; intra-class separation necessitates knowledge of high-level visual concepts, such as semantic differences, rather than lower-level patterns, such as shape or color. Moreover, when mixing images of the same class, latent features are naturally more similar, and their two-way loss functions will be updated in the same direction. Consequently, an intra-class mixture's latent features will encode more information that is more robust about a specific class than an inter-class mixture, where the two latent features may confuse or conflict with each other. Specifically, when the mixed representations have semantics that are more closely aligned, the information propagated into the two branches contains more information about the same class. Otherwise, when the mixed representations are from different classes, the disentangled features may not have semantics perfectly resembling their classes.

From these two observations, we introduce *semantics-enhanced sampling* for generating different percentages of same-classed mixtures r, a hyper-parameter we experimentally verify. After the model is trained by i-MAE using such kind of input data, we finetune the model with Mixup strategy (both baseline and our models) and cross-entropy loss. We use accuracy as the metric of semantics under this percentage of instance mixture:

$$\mathcal{M}_{\text{sem}} = -\sum_{i=1}^{n} \mathbf{t}_i \log\left(\mathbf{p}_i\right), \tag{7}$$

where $\mathbf{t}_i$ is the ground-truth label and $\mathbf{p}_i$ is the prediction. Since this sampling strategy will involve additional prior knowledge of same or different classes when sampling and mixing in a mini-batch, an alternative way to avoid using prior label information is clustering the samples to identify the same or different classes in a mini-batch, then conducting the sample assignment process.

4. Experiments and Analysis

In this section, we examine the effectiveness of the proposed i-MAE framework and analyze the properties of i-MAE's disentangled representations through empirical studies on an extensive range of datasets. First, we provide the details of datasets used and our implementation settings. Then, we thoroughly ablate our experiments, focusing on the linear and finetuning evaluations, as well as properties of linear separation, and semantic-enhanced mixture. Lastly, we illustrate the qualitative results and visualizations.

4.1. Datasets

CIFAR-10/100 [19] Both CIFAR datasets contain 60,000 tiny colored images sized 32×32. CIFAR-10 and 100 are split into 10 and 100 classes, respectively.
Tiny-ImageNet The Tiny-ImageNet is a scaled-down version of the standard ImageNet-1K consisting of 100,000 64x64 colored images, categorized into 200 classes.
ImageNet-1K [6] The ILSVRC 2012 ImageNet-1K classification dataset consists of 1.28 million training images and 50,000 validation images of 1000 classes.

4.2. Details of Implementation

Settings: We conduct experiments of i-MAE on CIFAR-10/100, Tiny-ImageNet, and ImageNet-1K. On CIFAR-10/100, we adjust MAE's structure to better fit the smaller datasets during unsupervised pre-training: ViT-Tiny [27] in the encoder and a lite-version of ViT-Tiny (4 layers) as the decoder. Our pre-training lasts 2,000 epochs with a learning rate 1.5×10^{-4} and 200 warm-up epochs. On Tiny-ImageNet, i-MAE's encoder is ViT-small and decoder is ViT-Tiny, trained for 1,000 epochs with a learning rate 1.5×10^{-4}. Additionally, we apply warm-up for the first 100 epochs, and use cosine learning rate decay with AdamW [18] optimizer as in vanilla MAE.

Unless otherwise stated, the default settings used in our experiments are a masking ratio of 75%, a mix factor sampled from a distribution $\beta(1.0, 1.0)$, and reconstructing both images with distillation loss for stronger representation.
Supervised Finetuning: In the finetuning process, we apply Mixup in all experiments to fit our pre-training scheme, and compare our results with baselines of the same configuration. On CIFAR-10/100, we finetune 100 epochs using the AdamW optimizer and a learning rate of 1.5×10^{-3}.
Linear Probing: For linear evaluation, we follow MAE [14] to train with no extra augmentations and use zero

weight decay. Similarly, we adopt an additional BatchNorm layer without affine transformation.

4.3. Main Results

In this section, we first provide the finetuning and linear evaluation results on various datasets. Following that, we empirically analyze our main findings: how separable are i-MAE embedded features and the advantage of semantics-enhanced sampling. Specifically, we quantitatively verify the linear separability of i-MAE's disentanglement. Then, we evaluate the performance improvement from utilizing the semantics-enhanced sampling strategy.

Finetuning and Linear Evaluation. Here we explore the performance gain from the architecture and training strategy adjustments. We evaluate our i-MAE's performance through finetuning and linear evaluation of regular inputs and targets. Finetuning and linear probing classification results are outlined in Tab. 1 and Tab. 2. It can be observed that i-MAE outperforms the baseline MAE and ViT from scratch by remarkable margins. As our features are learned from a harder scenario, they encode more information with more robust representation and classification accuracy. Besides, i-MAE shows a considerable performance boost with both evaluation methods.

Separability. Here, we show how i-MAE displays properties of linear separability, both visually and quantitatively, and demonstrate our advantage over vanilla MAE. We provide a visual comparison of the disentanglement capability in Fig. 3. In the first row, vanilla MAE does not perform well out-of-the-box when disentangling mixed inputs, with reconstructions representing the mixed input more so than the subordinate image. However, the latter two rows demonstrate that i-MAE performs reconstruction very well.

Since the mixture inputs of i-MAE is a linear combination of the two images, and our results show i-MAE's powerful ability to reconstruct both images, even at very low mixture ratios, we attribute such ability to i-MAE's disentanglement strongly correlating with the vanilla MAE's features. Now, we empirically illustrate the strength of the linear relationship between MAE's features and i-MAE's disentangled features with a linear separability metric. We employ different distances as our criteria, and results are reported in Tab. 4. Experimentally, we feed mixed inputs to i-MAE and a singular image to vanilla MAE, where the former produces disentangled features and the latter produces target features. Then, we train a single linear projection layer to fit the disentangled features to the target. *Fore* indicates that we directly calculate the metrics between the target features from vanilla MAE and our disentangled features. *Aft* indicates that we train the linear regressor's parameters to fit the target feature. *Baseline* is the model trained without the disentanglement module. It can be observed that our i-MAE has significantly smaller distances

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
from scratch	74.13	53.57	43.36
MAE	90.78	68.66	59.28
i-MAE	**92.00**	**69.50**	**61.63**

Table 1. Finetuning classification accuracy of ViT trained from scratch, baseline MAE, and the proposed i-MAE across different datasets.

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
MAE	72.47	32.57	19.62
i-MAE	**77.61**	**33.39**	**20.40**

Table 2. Linear probing accuracy of baseline MAE and proposed i-MAE across different datasets. *from scratch* is inapplicable here.

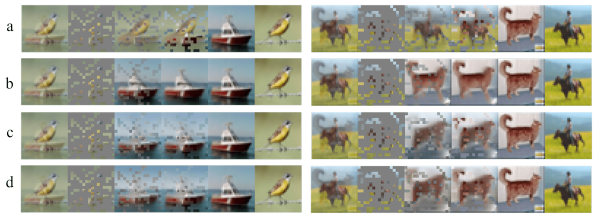

Figure 3. Qualitative ablation comparisons on CIFAR-10. **Row (a)**: baseline vanilla MAE; **(b)**: MAE with unmixed input; **(c)**: our i-MAE without distillation; and **(d)**: i-MAE with distillation.

than the vanilla model, indicating that such a scheme can obtain better linear separation ability.

Semantics. As shown in Tab. 3, we provide the ablation study of semantics-enhanced sampling strategy with intra-class mix rate r from 0.0 to 1.0. In the table, *baseline* represents the vanilla MAE model and 0 is the regular sampling strategy without semantics-enhanced sampling. We can observe that the performacne has a certain increase when the same-class ratio is employed with 0.5 or 1.0.

As discussed in the Sec. 3.2, we emphasize that our enhanced performance comes from i-MAE's ability to learn more separable features with the disentanglement module, and the enhanced semantics learned from training with *semantics-enhanced sampling*. Our classification results show the cruciality of MAE learning features that are linearly separable, which can help identify between different classes. However, to correctly identify features with their corresponding classes, semantically rich features are needed, which can be enhanced by the intra-class mix sampling strategy.

4.4. Ablation Study

In this section, we perform ablation studies on i-MAE to concretely examine the property of linear separability and its existence at different mix-levels. Then, we analyze the effects of *semantics-enhanced mixture* on i-MAE learned representations.

Ablation for Linear Separability. In this work, motivated by approaches measuring latent disentanglement with linear means, we propose mixture strategies and semantic mixing

Same-class Ratio	CIFAR-10		CIFAR-100		Tiny-ImageNet	
	Finetune	Linear	Finetune	Linear	Finetune	Linear
MAE (baseline)	90.78	72.47	68.66	32.57	59.28	19.62
0.0	91.67	70.53	68.34	29.22	60.91	18.23
0.5	**92.34**	72.80	**69.50**	30.11	60.58	18.51
1.0	91.60	**77.61**	69.33	**33.39**	**61.13**	**20.40**

Table 3. Ablation study of semantics-enhanced sampling strategy with intra-class mix rate r from 0.0 to 1.0 (*baseline* is the vanilla MAE, 0.0 indicates the regular sampling strategy). The lower bound represents that inputs are mixes with all different classes, $r = 1.0$ indicates the model is pre-trained with solely mixtures of same-labeled instances.

Score	Method	CIFAR-10		CIFAR-100		Tiny-ImageNet		ImageNet-1K	
		Fore	Aft	Fore	Aft	Fore	Aft	Fore	Aft
NRMSE ↓	MAE (Baseline)	0.247	0.223	0.254	0.208	**0.701**	0.624	**0.639**	0.334
	i-MAE w/o distill	0.372	0.221	0.368	0.202	1.023	0.627	0.850	0.313
	i-MAE	**0.179**	**0.188**	**0.176**	**0.181**	0.927	**0.597**	0.869	**0.299**
R^2 ↑	MAE (Baseline)	0.336	0.438	0.242	0.482	0.269	0.429	-	0.625
	i-MAE w/o distill	-	0.540	-	0.513	-	0.420	-	0.714
	i-MAE	**0.807**	**0.637**	**0.633**	**0.609**	**0.275**	**0.480**	-	**0.736**
Cos ↑	MAE (Baseline)	0.672	0.665	0.626	0.694	**0.643**	0.662	**0.501**	0.768
	i-MAE w/o distill	0.062	0.741	0.007	0.718	-	0.655	0.002	0.816
	i-MAE	**0.807**	**0.796**	**0.794**	**0.779**	0.000	**0.696**	0.009	**0.837**

Table 4. Linear separation metric using NRMSE, R^2 (R-Squared or coefficient of determination) and *cosine*-similarity as measurements for patch-wise feature comparisons, calculated before and after linear regression on CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet. Reported results are from linear regressor's prediction on validation set. *i-MAE* and *i-MAE without distillation* are embedding after disentanglement. ↓ indicates lower is better and ↑ indicates higher is better.

Figure 4. Comparisons between mask ratios on Tiny-ImageNet validation set. i-MAE produces enhanced visual reconstructions from lower masking ratios when reconstructing images.

for the purpose of learning stronger and more linearly separable features. To begin, we thoroughly perform our ablation experiments on a diverse group of datasets (ImageNet-1K is performed for final evaluation) and demonstrate how i-MAE's learned features display linear separability with different settings. Specifically, we experiment with the separability of the following aspects of our methods: (1) constant and probability mix factors; (2) masking ratio of input mixtures; (3) different ViT architectures. As previously mentioned, since our model produces two reconstructions, our visual results demonstrate the subordinate branch, which is worse in quality, to demonstrate the effectiveness of our method.

(1) Mix Ratio. To demonstrate the separable nature of the input mixtures for reconstructions, we compared different mixture factors between 0 and 0.5, and random mixture ratios from a Beta distribution. Intuitively, lower mixing ratios contain less meaning information that the encoder may easily confuse with noise, whereas higher ratios destroy the subordinate-dominant relationship. Experimentally, we observe matching visual results shown in the supplementary materials The better separation performance near 0.3 indicates that i-MAE features are better dichotomized when the mix factor is balanced between noise and useful signals. Whereas below 0.15, the subordinate image is noisy and reconstructions are not interpretable, mixing ratios above 0.45 break the subordinate relationship between the two images, and the two features are harder to distinguish from each other. Moreover, Fig. 1 presents a problematic case where a mix factor of 0.45 reconstructs dominant features (hence the green patches in the background).

(2) Mask Ratio. In i-MAE, visible information of the subordinate image is inherently limited by the unbalanced mix ratio, in addition to masking. Hence, a high masking ratio (75% [14]) may not be necessary to suppress the amount of information the encoder sees, and we consequently attempt lower ratios of 50%, 60% to introduce more information about the subordinate target. As shown in Fig. 4, a lower masking ratio of 0.5 or 0.6 can significantly improve reconstruction quality.

Combining our findings in mix and mask ratios, we empirically find that i-MAE can compensate for the informa-

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
i-MAE-Sub	74.23	49.83	49.50
i-MAE	**92.00**	**69.50**	**61.63**

Table 5. Single and dual reconstruction ablation. We compare i-MAE pre-trained to only reconstruct the subordinate image (i-MAE-sub) and to reconstruct both (i-MAE). Better representations are learned with dual reconstructions.

tion loss at low ratios with the additional alleviation of more visible patches (lower mask ratio). Illustrated in Fig. 1, we display a case of i-MAE's reconstruction succeeding in separating the features an input with $\alpha = 0.1$ mix factor and 0.5 masking ratio. Through studying the mix ratio and masking ratio, we reveal that i-MAE can learn linearly separable features under two conditions: **(i)** enough information about both images must be present (determined by the trade-off between mask ratio and mix ratio). **(ii)** the image-level distinguishing relationship between minority and majority (determined by mix ratio) is potent enough for i-MAE to encode the two images separately.

(3) ViT Backbone Architecture. We studied whether different scales of ViT effect linear separation in Appendix. Our results show that larger backbones are not necessary for i-MAE to disentangle features on small datasets, as the insufficient training data cannot fully utilize the capability. However, large ViTs are crucial to the large-scale ImageNet-1K dataset.

Ablation for Degree of Semantics. We provide the ablation study on different ratios of mixing the same class samples within a mini-batch.

Semantic Mixes. Depending on the number of classes and their overall size, datasets in pristine states usually contain around 10% (e.g., CIFAR-10) to <1% (e.g. ImageNet-1K) samples pertaining to the same class, meaning that by default, uniformly random sampling mixtures will most likely be of different objects. On the other hand, the *semantics-enhanced mixture* scheme examines whether the introduction of semantically homogeneous mixtures affects the classification performance. That is, we intentionally test to see if similar instances during pre-training negatively influence the classification performance.

As shown in Tab. 3, after i-MAE pre-training, we perform finetuning and linear probing on classification tasks to evaluate the degree of semantics learned given different amounts of intra-class mix r. From Tab. 3, we discover that i-MAE overall has a stronger performance in finetuning and linear probing with a non-zero same-class ratio. Specifically, a high r of 1.0 increases the accuracy in linear evaluation most in all datasets, meaning that the quality of learned features is best and separated, and it gains a strong prior of category information for semantically enhanced mixtures. On the other hand, setting $r = 0.5$ is advantageous during finetuning, as it gains a balanced prior of separating both intra- and inter-class mixtures.

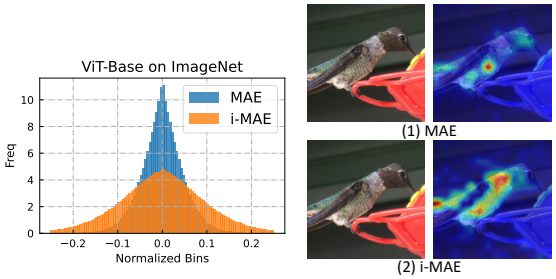

Figure 5. Left is the comparison of weight distribution between MAE and i-MAE pre-trained on ImageNet-1K. Our weights are sparser. Right is the comparison of attention maps. Our model has a wider view on the input image which encodes more information.

Single vs. Dual Reconstruction

To verify the design in Sec. 3.1.1, we perform finetuning with i-MAE pre-trained on linear disentanglement with single and dual reconstructions. The results are shown in Table 5. It is clear that dual reconstructions achieve better representational performance.

4.5. Visualizations

We provide the visualization comparison of weight distributions and attention mappings. In Fig. 5 (left), we show the difference of weight distributions between MAE and i-MAE on ImageNet-1K dataset. Compared with vanilla MAE's weights, our model's weights have a better diversification distribution, indicating that i-MAE forces the model to spread out and incorporate more patterns over the entire image mix range. In Fig. 5 (right), our model has a wider focusing area, which also indicates that more information is encoded in the trained model through the proposed mixture training scheme.

5. Conclusion

It is nontrivial to understand why Masked Image Modeling (MIM) in the self-supervised scheme can learn useful representations for downstream tasks without labels. In this work, we have introduced a novel interactive framework upon Masked Autoencoders (i-MAE) to explore two critical properties in latent features: *linear separability* and *degree of semantics*. We identified that the two specialties are the core for superior latent representations and revealed the reasons where is the good transferability of MAE from. Moreover, we proposed two metrics to evaluate these two specialties quantitatively. Extensive experiments are conducted on CIFAR-10/100, Tiny-ImageNet, and ImageNet-1K datasets to demonstrate our discoveries and observations in this work. We also provided sufficient qualitative results and analyses of different hyperparameters. We hope this work can inspire more studies on the understanding and improvement of the MIM frameworks for the self-supervised pretraining in the future.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 1, 2

[3] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. 3, 4

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. 2

[8] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 1

[9] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018. 4

[10] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 2

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 1

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 513–520, Madison, WI, USA, 2011. Omnipress. 4

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2, 3, 5, 7

[15] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 1

[16] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018. 4

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 4

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 5

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 5

[20] Jihao Liu, Xin Huang, Osamu Yoshie, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning, 2022. 2

[21] Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders, 2018. 3

[22] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *CoRR*, abs/1912.01991, 2019. 3

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[24] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 2

[25] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2216–2224, 2022. 2

[26] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *CoRR*, abs/1503.02406, 2015. 4

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 5

[28] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 3

[32] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1

[33] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 2