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

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

Appendix

In the appendix, we provide the detailed configurations of our experiments and elaborate with more visualizations that are supplementary for our main text, specifically:

• Section 1 "Implementation Details": implementation details and configuration settings for unsupervised pre-training and supervised classification.

• Section 2 "More Visualizations": additional reconstruction examples on various datasets.

• Section 3 "Pseudocode": a PyTorch-like pseudocode for our detailed procedure of the proposed framework.

1. Implementation Details in Self-supervised Pre-training, Finetuning, and Linear Evaluation

ViT architecture. In our non-ImageNet datasets, we adopt smaller ViT backbones that generally follow [2]. The central implementation of linear separation happens between the MAE encoder and decoder, with a linear projection layer for each branch of reconstruction. A shared decoder is used to reconstruct both images. A qualitative evaluation of different ViT sizes on Tiny-ImageNet is displayed in Fig. 1. The perceptive difference is not large, and generally, ViT-small/tiny are sufficient for non-ImageNet datasets.

Pre-training. The default setting for pre-training is listed in Tab. 1. On ImageNet-1K, we strictly use MAE's specifications. For better classification performance, we use normalized pixels [1] and a masking ratio of 0.75. For better visual reconstructions, we use a lower masking ratio of 0.5 without normalizing target pixels. In CIFAR-10/100, and Tiny-ImageNet, reconstruct ordinary pixels.

Semantics-enhanced sampling. The default settings for our semantics-enhanced mixtures are listed in Tab. 2. We modified the dataloader to mix, within a mini-batch, $\boldsymbol{r}$ percent of samples that have homogenous classes and $1-\boldsymbol{r}$ percent that have different ones.

Classification. For the classification task, we provide the detailed settings of our finetuning process in Tab. 3 and linear evaluation process in Tab. 4.

Algorithm 1: PyTorch-style pseudocode for dual reconstruction targets on i-MAE.

```
# α:  mixture ratio alpha
# b = hyperparameter for the Beta
 Distribution
# c = coefficient for balancing
 distillation loss and reconstruction loss
# E_i, E_m:  i-MAE encoder, vanilla encoder
 for distillation
# D_i:  i-MAE decoder
# F_1, F_2:  linear decomposition layer 1, 2
def forward(img):
    a = random.beta(b, b)
    x1, x2 = img, perm(img) # Perm can be
     random permutation for inner batch mix,
     or semantics-enhanced sampling
    mix = α * x1 + (1-α) * x2
    # subordinate and dominant image
    x1, x2 = x1, x2 if α < 0.5 else x2, x1
    w1, w2 = E_m(x1), E_m(x2)
    z1, z2 = F_1(E_i(mix)), F_2(E_i(mix))

    loss = c * distill(w1,w2,z1,z2) +
     recon(z1, z2, x1, x2)
    return loss

def distill(w1, w2, z1, z2):
    loss = l2_loss(z1, x1)
    loss += l2_loss(z2, x2)
    return loss

# D_i:  i-MAE decoder
# norm_pix:  normalization function on pixel
 of each masked patch as the target
def recon(z1, z2, x1, x2):
    t1, t2 = norm_pix(x1), norm_pix(x2)
    p1, p2 = D_i(z1), D_i(z2)
    loss = MSE(p1, t1) + MSE(p2, t2)
    return loss
```

2. More Visualizations

We provide extra examples of pre-trained i-MAE reconstructing only the subordinate image. Fig. 2 are visualizations on CIFAR-100 at mix ratios from 0.1 to 0.45, in 0.05

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

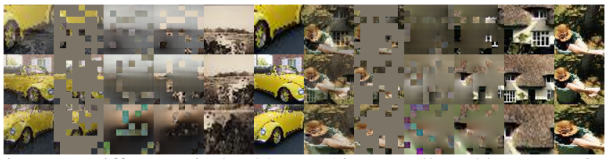

Figure 1. Different ViT backbones (tiny, small, and base) on Tiny-ImageNet. Reconstruction quality is moderately improved when a larger backbone is used.

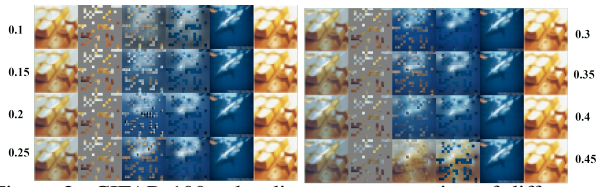

Figure 2. CIFAR-100 subordinate reconstruction of different ratios marked on the left and right side. Similarly, reconstructions at 0.45 are confused with the dominant image.

Figure 3. Uncurated reconstructions of CIFAR-100 validation images using *semantics-enhanced mixture* from 0.0 (topmost) to 1.0 (bottom), in 0.1 intervals.

steps. As depicted in Fig. 6 and Fig. 7, we produce finer ranges of reconstructions from 0.05 to 0.45. In most cases, mixture rates above 0.4 tend to show features of the dominant image. This observation demonstrates that a low mixture rate can better embed important information separating the subordinate image.

3. PyTorch Styled Pseudocode

The pseudocode of our mixture and subordinate reconstruction approach is shown in Algorithm 1. In our full-fledged

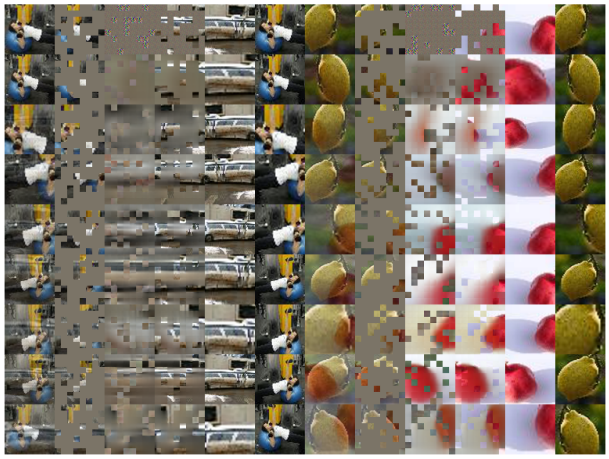

Figure 4. Uncurated Tiny-ImageNet reconstructions of different mix ratio, from 0.05 to 0.45, subordinate images.

Figure 5. Visual reconstructions of Tiny-ImageNet validation images using *semantics-enhanced mixture* pre-trained i-MAE.

i-MAE, we employ two distillation losses for two linear separation branches. "x2" is sampled by a permutation within a mini-batch. Alternatively, we can also employ the semantics-enhanced sampling scheme for it to create r percent of samples from the same class.

References

[1] 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] 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. 1

Config	CIFAR-10/100	Tiny-ImageNet	ImageNet-1K
base learning rate	1.5e-4	1.5e-4	1e-3
batch size	4,096	4,096	4,096
Mask Ratio	0.75	0.75	0.5
optimizer	AdamW	AdamW	AdamW
optimizer momentum	0.9, 0.95	0.9, 0.95	0.9, 0.95
augmentation	None	RandomResizedCrop	RandomResizedCrop

Table 1. Self-supervised pre-training configurations on CIFAR-10/100, Tiny-ImageNet, ImageNet-1K. For better visualizations, we use a 0.5 mask ratio on ImageNet.

Config	CIFAR-10/100	Tiny-ImageNet
Object mix range	0.0, 0.25, 0.5, 1.0	0.0, 0.25, 0.5, 1.0
Image mix ratio	Beta$(1.0, 1.0)$	Beta$(1.0, 1.0)$
base learning rate	1.5e-4	3.5e-4
batch size	4,096	4,096
Mask Ratio	0.75	0.75
optimizer	AdamW	AdamW
optimizer momentum	0.9, 0.95	0.9, 0.95
augmentation	None	RandomResizedCrop

Table 2. Pre-training Configurations of with the *semantics-enhanced mixture* scheme.

Config	CIFAR-10/100	Tiny-ImageNet	ImageNet-1K
Object mix range	0.0 - 1.0	0.0 - 1.0	0.0, 0.25, 0.5, 1.0
Image mix ratio	Beta$(1.0, 1.0)$	Beta$(1.0, 1.0)$	0.8
base learning rate	1e-3	1e-3	1e-3
batch size	128	256	1,024
epochs	100	100	25
optimizer	AdamW	AdamW	AdamW
optimizer momentum	0.9, 0.999	0.9, 0.999	0.9, 0.999
augmentation	Mixup	Mixup, RandomResizedCrop	Mixup, RandomResizedCrop

Table 3. Finetune Classification Configurations.

Config	CIFAR-10/100	Tiny-ImageNet	ImageNet-1K
Object mix range	0.0 - 1.0	0.0 - 1.0	0.0, 0.25, 0.5, 1.0
Image mix ratio	0.35	0.35	0.35
base learning rate	1e-2	1e-2	1e-2
batch size	128	256	1,024
epochs	200	200	25
optimizer	SGD	SGD	SGD
optimizer momentum	0.9, 0.999	0.9, 0.999	0.9, 0.999
augmentation	None	RandomResizedCrop	RandomResizedCrop

Table 4. Linear Classification Configurations.

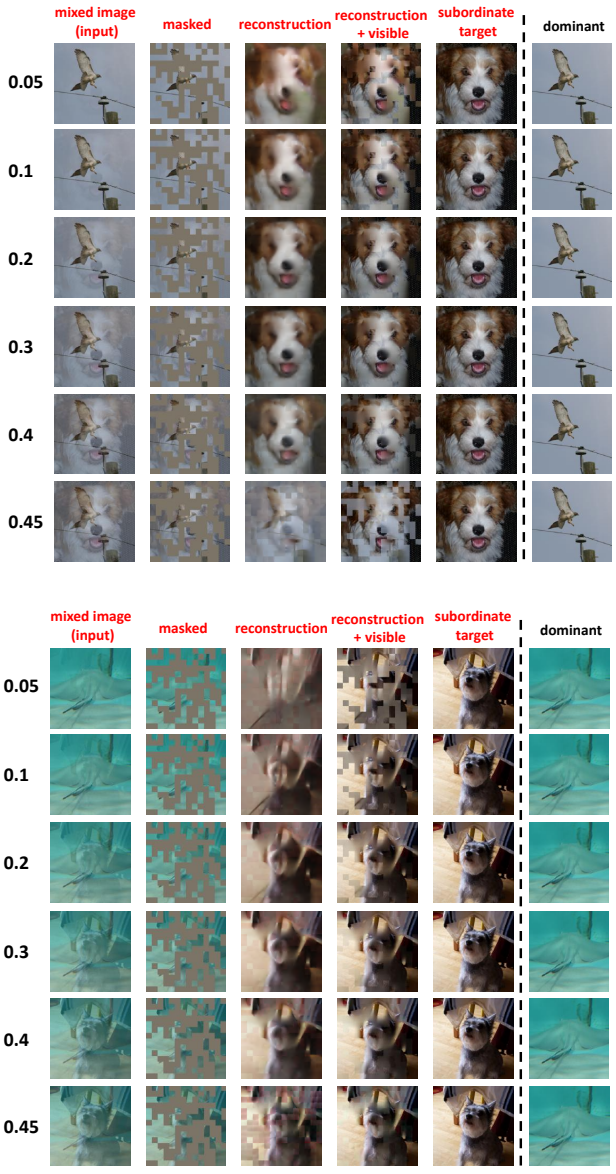

Figure 6. More reconstructions results of i-MAE on ImageNet-1K *validation* images with different mixing coefficients α (listed on the left), 0.5 mask ratio, and with distillation. Visual results are the subordinate reconstructions.

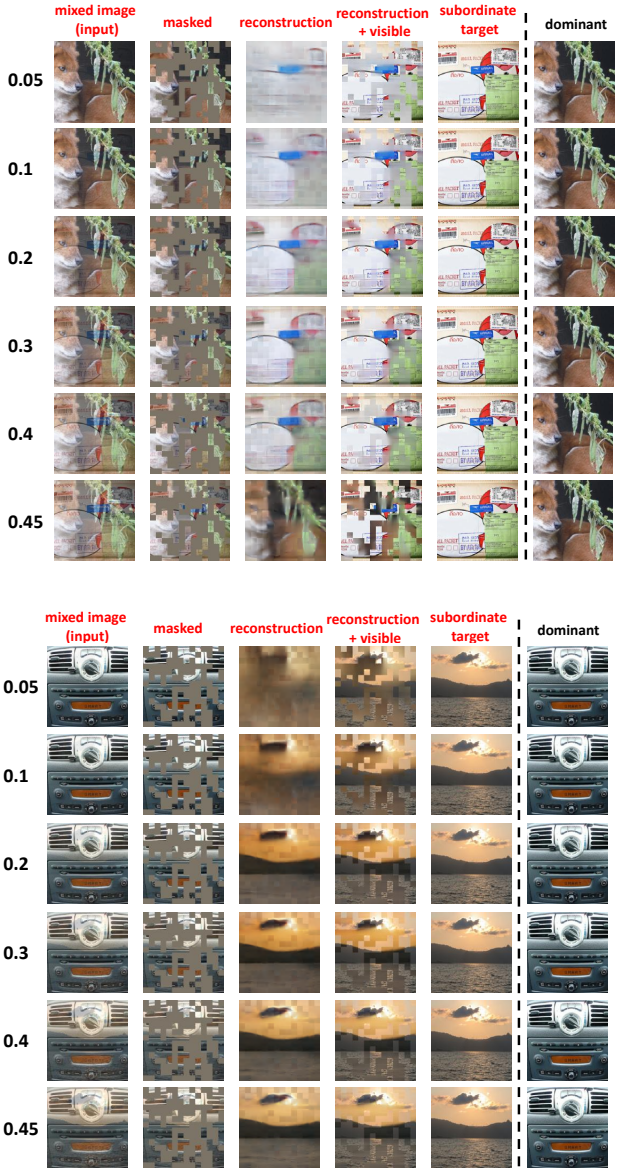

Figure 7. More reconstructions results of i-MAE on ImageNet-1K *validation* images with different mixing coefficients α (listed on the left), 0.5 mask ratio, and with distillation loss. Visual results are the subordinate reconstructions.