Understanding Masked Image Modeling via Learning Occlusion Invariant Feature

Xiangwen Kong[1], Xiangyu Zhang[1,2]
[1] MEGVII Technology, [2] Beijing Academy of Artificial Intelligence
{kongxiangwen,zhangxiangyu}@megvii.com

A. Settings of Augmentation Experiments

Augmentation strategies. Generally in MAE, the *source image* which is send into encoder and the *target image* which is the target to reconstruct are always the same. Here we try to add additional augmentations on the source image. After the random-resize-cropped and random horizontal-flip augmentation, the image is cropped to 224×224. Then we try the compositions of three additional augmentation strategies on the source image. The definitions of augmentation strategies are described as follow:

1. Patch masking (Fig. 1(b)): we divide the image into non-overlapping 16×16 patches and randomly mask 75% patches, then the image is occluded by small and neat black blocks.

2. Cutmix (Fig. 1(c)): we cropped a patch from the original image then resize the patch and paste it on the image. The image is occluded by an object rather than small blocks.

3. Color augmentation (Fig. 1(d)): we use the same color augmentation as SimSiam [1]. Although there is no occlusion, the entire source image and target image are slightly different in color and texture after color augmentation.

Model configurations. We use ViT-S/16 as backbone, and the models are trained and evaluated on ImageNet-100[1] classification. For C-MAE, we use normalized L2 loss [2] as measurement. Other configurations are the same as default.

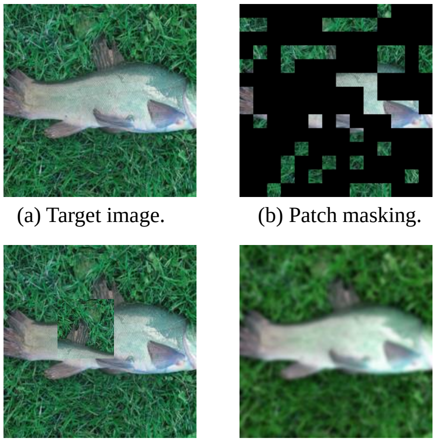
(a) Target image.
(b) Patch masking.
(c) CutMix.
(d) Color augmentation.

Figure 1. Visualization of different augmentation strategies.

B. More Visualization Experiments

B.1. Occlusion-invariance of Few Images Pretrained MAE

Here we discuss the occlusion invariance of a few images pretrained MAE models. We use **CKA similarities** between the representations generated by the masked image and the full image under different mask ratios as protocol. The numbers (0.1 to 0.9) indicate the mask ratios (i.e. percentages of image patches to be dropped) of the test images respectively. The higher CKA similarity with a large mask ratio means the model learns better occlusion invariance.

Fig. 2 shows the CKA similarities of MAE pretrained with different amounts of data. As the figures show, the model learns occlusion invariance even pretrained with one image. Unfortunately, the model does not keep the occlusion invariance after finetuning. When the mask ratio increase to 0.7, the CKA similarities drop significantly below 0.5. In Comparison, full-set pretrained MAE is not so sensitive to the change of mask ratio (after 0.7) after finetuning.

Furthermore, we discuss the relationship between occlusion invariance with overfitting. We train the MAE with different training epochs on 10 images and plot the CKA similarities. Results in Fig. 3 show that, overfitting affects the learning of occlusion invariance, and causes the performance to drop. We further explore the way to prevent overfitting, *using stronger data augmentation*, whether beneficial to maintain occlusion invariance. As shown in Fig. 3, even the finetuning results increase a little when using stronger augmentations, the occlusion invariance does not been improved.

[1] https://www.kaggle.com/datasets/ambityga/imagenet100

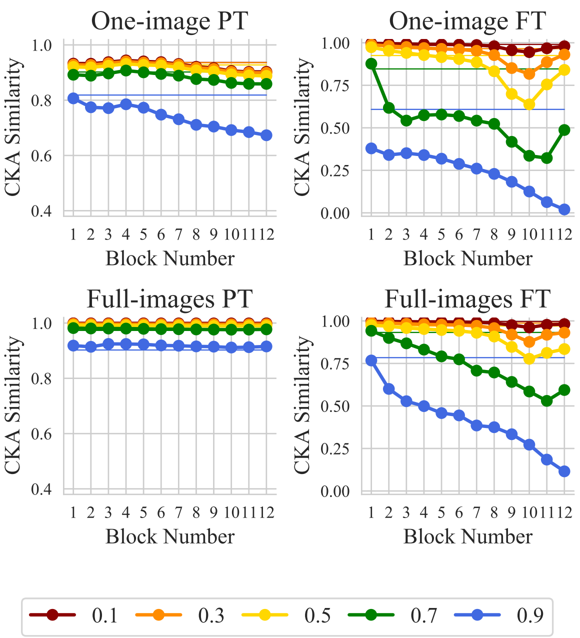

Figure 2. CKA similarity between the representations generated by the masked image and the full image respectively under different mask ratios. **One-image PT** indicates the model that pretrained on one image for 5 epochs, and **Full-images PT** indicates the model pretrained on ImageNet training set for 100 epochs. The finetuning models (**FT**) are all trained on ImageNet training set for 100 epochs.

B.2. Comparison with Human Recognition

[3] shows that, ViT behaves more like humans in classification, and we wonder whether our proposed siamese framework learns more high-level perception. Following the method in [3], we plot the shape bias of MIM models in Fig. 4.

Fig. 4 shows the shape bias of MAE, MoCov3, R-MAE and C-MAE. As shown in the figure, the grey line represents the supervised trained model, which has the lowest shape bias. That means fully supervised learning prefers to learn texture information rather than self-supervised pretrained models. Both MAE (blue line) and R-MAE (green line) learn less shape bias than MoCo (yellow line) and C-MAE (orange line). We speculate that it is because the target of the pretext task of MIM is closer to the original images (or exactly the origin images), which makes the model learn more texture features. Additionally, C-MAE learns a similar shape-bias compared with MoCo v3. The results indicate that instance-wise learning is not necessary for models to learn as human does, learning occlusion invariance could also improve the ability of the model to learn shape-bias. When training longer, all masked-based models are biased to learn texture features. We conclude that the masked-based models could learn the ability to complete object shape quickly in a few epochs, and then learn to reconstruct the texture of images.

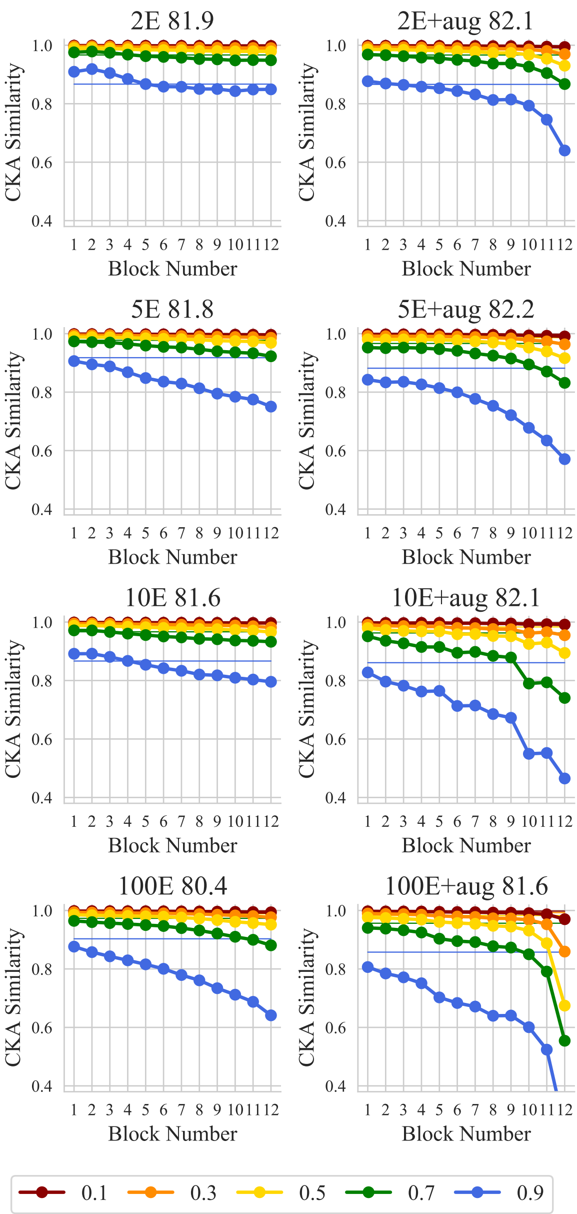

Figure 3. CKA similarity between the representations generated by the masked image and the full image respectively under different mask ratios. The number on the subtitle is the finetuning result of the model. All models are pretrained on 10 images for different epochs and finetuned on full ImageNet training set for 100 epochs. N**E** means the model pretrained for N epochs. **+aug** means adding stronger augmentations.

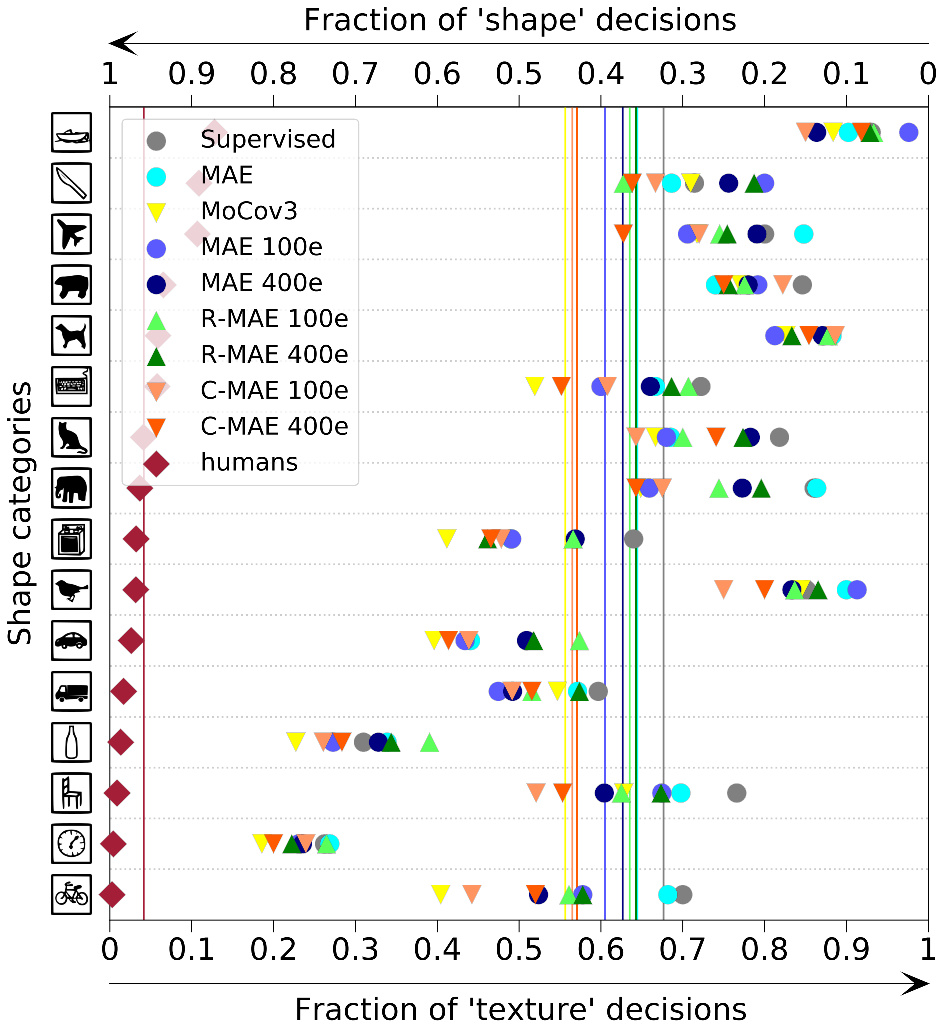

Figure 4. Shape bias of MAE, MoCov3, R-MAE and C-MAE pretrained on ImageNet. The vertical line is the average shape bias of 16 classes.

References

[1] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 1

[2] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[3] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? *CoRR*, abs/2105.07197, 2021. 2