Supplementary Material

Qiu Han[1], Gongjie Zhang[1,2], Jiaxing Huang[1], Peng Gao[3], Zhang Wei[3], Shijian Lu[1*]

[1]S-Lab, Nanyang Technological University
[2]Black Sesame Technologies
[3]Shanghai Artificial Intelligence Laboratory

han023@e.ntu.edu.sg, Gjz@ieee.org, {Jiaxing.Huang, Shijian.Lu}@ntu.edu.sg, gaopeng@pjlab.org.cn

A. Implementation Details

A.1. Semantic Segmentation in ADE20K

We use UperNet [4] and adopt the same model structure as in [2]. We follow the implementation of MMSeg [1] and search for some of the hyper-parameters. As a result, we use a learning rate of 1.4e-4, weight decay of 0.03, and layer-wise decay rate of 0.75, while the original settings are 1e-4, 0.05, and 0.65 respectively. With the new parameter, our Efficient MAE can perform on par with the original MAE for semantic segmentation in ADE20K dataset.

A.2. Object Detection and Segmentation in COCO

We adopt ViTDet [3] with the ViT-Base model as our baseline. However, due to the limitation of computing resources, we are unable to re-implement the experiments based on their original setting. We thus train the ViTDet for 25 epochs with a batch size of 16 (the original setting is 100 epochs with a batch size of 64). The baseline of the original MAE achieves 50.1 AP_{bbox} / 44.7 AP_{mask}. For Efficient MAE, we also search for the hyper-parameters and find that a slightly higher layer-wise decay rate can lead to higher performance. We thus use a layer-wise decay rate of 0.8 instead of 0.7 in the ViTDet for the original MAE ViT-Base. This is consistent with the hyper-parameter in ADE20K. We therefore speculate that it is possible that a larger mask ratio during pre-training would lead to a preference for a higher layer-wise decay rate on downstream tasks.

B. Trade-off of Reconstruction Difficulty

In SimMIM [5], a metric AvgDist is proposed to measure the overall difficulty and effectiveness of masked image modeling (MIM). It came to the conclusion that the difficulty of reconstruction is better to be moderate. Reconstruction that is too easy or too hard might hurt the performance. This seems controversial to our methods that apply higher weight on easy reconstruction targets. However, we argue that though implemented in different ways, our method is consistent with the conclusions obtained by AvgDist in SimMIM [5]. On the one hand, though we apply higher weight for easy patches, the reconstruction task is still of moderate difficulty since there are few easy patches when the mask ratio is extremely high as shown in Fig.2. This is different from SimMIM where the decrease of mask ratio might lead to an oversimple reconstruction task, while our method mainly focuses the loss optimization and has little effect on the total task difficulty. On the other hand, by applying lower weight and directly discarding the extremely hard targets in Decoder Masking, we adjust the overall reconstruction difficulty of MIM under an extremely high mask ratio into the moderate level which is consistent with the moderate AvgDist in SimMIM. The simplified reconstruction task is able to mitigate the severe performance drop under extremely high mask ratios, thus enabling the per-training with higher efficiency.

C. Definitions of Different Metrics in Tab.5

The metrics presented in Tab.5 are specifically designed to measure the reconstruction difficulty. The first metric, "Number of Patches in 3×3", computes the density (number) of visible patches within a 3×3 neighborhood around each masked patch. In the case of the "Sum of Patch Distance in 5×5", we expand the neighborhood range into 5×5, and weighted sum the number based on the distance from the visible patches to the current masked patch. We assign lower weights to distant visible patches, equaled to the inverse of distances, as they contribute loss to the reconstruction. The last metric, "Nearest Patch Distance" is similar to our proposed $\mathrm{P}^2\mathrm{Dist}$. However, it solely measures the distance between the center of each masked patch and the center of its nearest visible patch. Consequently, this metric lacks the capability to effectively model complex masked patterns.

*Corresponding author.

References

[1] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[3] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022. 1

[4] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 1

[5] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, 2022. 1