Improving Pixel-based MIM by Reducing Wasted Modeling Capability

Yuan Liu[1], **Songyang Zhang**[1,‡], **Jiacheng Chen**[2], **Zhaohui Yu**[3], **Kai Chen**[1,‡]
Dahua Lin[1,3]
[1]Shanghai AI Laboratory [2]Simon Fraser University [3]The Chinese University of Hong Kong
‡ Corresponding author

A. Appendix

A.1. Pre-training

The settings for pre-training strictly follows those in MAE[8] and PixMIM[13], with details shown below:

config	value
optimizer	AdamW [15]
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.95 [3]
batch size	4096
learning rate schedule	cosine decay [14]
warmup epochs [7]	40

Table 1: Pre-training setting of MFF_{MAE} and $\text{MFF}_{\text{PixMIM}}$

A.2. Fine-tuning and linear probing

We also stick to the settings in MAE[8] for the ViT-B[5] model concerning fine-tuning and linear probing. Since our objective is to measure the enhancement brought by MFF and not attain the state-of-the-art (SOTA) performance, we employ the same settings as ViT-B without any specific adjustments for ViT-S.

config	value
optimizer	LARS [18]
base learning rate	0.1
weight decay	0
optimizer momentum	0.9
batch size	16384
learning rate schedule	cosine decay
warmup epochs	10
training epochs	90
augmentation	RandomResizedCrop

Table 2: Linear probing setting of MFF $_{\text{MAE}}$ and MFF $_{\text{PixMIM}}$.

A.3. Object detection and segmentation in COCO

All these settings also strictly follow those in MAE[8] but choose the commonly used 2× settings, which fine-tunes the model on COCO[12] for 25 epochs.

config	value
optimizer	AdamW[15]
base learning rate	1e-3
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.999
layer-wise lr decay [4, 1]	0.75
batch size	1024
learning rate schedule	cosine decay
warmup epochs	5
training epochs	100
augmentation	RandAug (9, 0.5) [9]
label smoothing [17]	0.1
mixup [20]	0.8
cutmix [19]	1.0
drop path [11]	0.1

Table 3: End-to-end fine-tuning setting of MFF $_{\text{MAE}}$, MFF $_{\text{PixMIM}}$

A.4. Semantic segmentation in ADE20K

We stick to the settings used in MAE[8] and PixMIM[13], fine-tuning the pre-trained model end-to-end for 16k iterations with a batch size of 16.

A.5. Selected indices of the ablation study

Inspired by the results of the pilot experiment depicted in Figure 1 of the main paper, we choose layer_0 as the shallow layer, and layer_{10} as the deep layer for the ablation experiment outlined in Table 3(a). Additionally, for ablation study in Table 3(b), we have selected additional two, four, and ten layers, evenly distributed between layer_0 and the output layer (layer_{11}). The detailed indices for Table 3(b) is shown in the Table 4.

In addition, similar to the pilot experiment in Figure 1 of the main paper, we observe the weight for each layer of all experiments in Table 3(b) of the main paper. Just as shown in Figure 1, no matter in which case, the model increasing relies on these shallow layers for the reconstruction tasks, indicating the significance of injecting low-level information into the output layer.

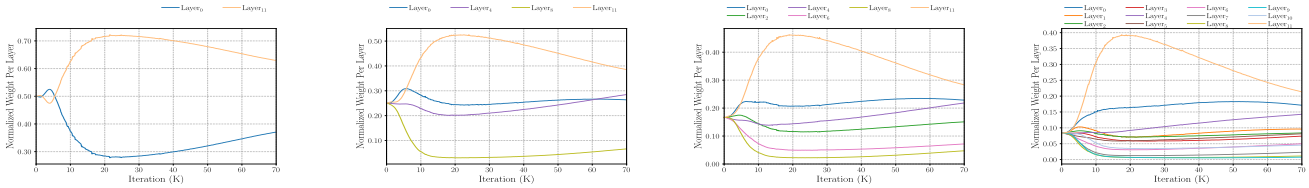

Figure 1: **Model increasingly relies on shallow layers.**

num layers	indices
1	11
2	0,11
4	0,4,8,11
6	0,2,4,6,8,11
12	0,1,2,3,4,5,6,7,8,9,10,11

Table 4: **Detailed indices for Table 3(b) of the main paper.** We try to make the additionally selected indices are evenly distributed between the first layer and last layer.

A.6. Transfer learning

We also study transfer learning where we pre-train on ImageNet-1K and fine-tune on several smaller datasets. We follow the training recipe and protocol in DINO[2]. MFF $_{\text{MAE}}$ consistently outperforms MAE on CIFAR10, CIFAR100, and Stanford Cars. As shown in the following table, MFF $_{\text{MAE}}$ consistently improves MAE on all datasets.

Method	Epoch	CIFAR10	CIFAR100	Cars
MAE	800	98.4	89.4	94.3
MFF $_{\text{MAE}}$	800	98.6 (+0.2)	90.3 (+0.9)	94.7 (+0.4)

Table 5: **Transfer learning on smaller datasets.**

A.7. Feature-based MIM does not Suffer from being Biased toward Low-level Feature

To supplement the findings in Figure 6 of the main paper, we apply multi-level feature fusion (MFF) to EVA[6] and MILAN[10], and evaluate their performance with linear probing, fine-tuning and semantic segmentation. Detailed results are shown below:

Method	Epoch	lin	seg	ft
EVA	400	69.0	49.5	83.7
MFF$_{\text{EVA}}$	400	68.9	49.4	83.8
MILAN	400	79.9	52.7	85.4
MFF$_{\text{MILAN}}$	400	79.7	52.9	85.0

As shown in the table above, MFF brings marginal improvements to feature-based MIMs, consistent with the findings in Figure 6 of the main paper.

A.8. The Effect of Deep Supervision

To exclude the influence of deep supervision[16], we detach all shallow layers before fusing with the last layer (MFF$^{\text{detach}}_{\text{MAE}}$), ensuring that gradients do not propagate through these shortcuts to the shallow layers. As shown in the table below, deep supervision alone does not improve MAE, and MFF's improvements come from alleviating the problem of being biased toward high-freq components.

Method	model	epoch	lin	seg	ft
MFF$_{\text{MAE}}$	ViT-B	800	67.0	47.9	83.6
MFF$^{\text{detach}}_{\text{MAE}}$	ViT-B	800	66.8 (-0.2)	48.0 (+0.1)	83.5(-0.1)

References

[1] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *ArXiv*, 2021. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021. 2

[3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 1

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020. 1

[5] 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. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual represen-

tation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[7] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *ArXiv*, abs/1706.02677, 2017. 1

[8] 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 (CVPR)*, 2022. 1

[9] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: better training with larger batches. *ArXiv*, abs/1901.09335, 2019. 1

[10] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and S. Y. Kung. Milan: Masked image pretraining on language assisted representation. *ArXiv*, abs/2208.06049, 2022. 2

[11] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 646–661, Cham, 2016. Springer International Publishing. 1

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[13] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*, 2023. 1

[14] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 1

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[16] Sucheng Ren, Fangyun Wei, Samuel Albanie, Zheng Zhang, and Han Hu. Deepmim: Deep supervision for masked image modeling. *arXiv preprint arXiv:2303.08817*, 2023. 2

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 1

[18] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv: Computer Vision and Pattern Recognition*, 2017. 1

[19] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. 1

[20] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1