Supplemental Material of Stare at What You See: Masked Image Modeling without Reconstruction

A. Linear Probing

We also perform linear probing by appending a linear classifier after the final layer of the pre-trained model, following MAE [6]. We use a mini-batch size of 16384, and an initial base learning rate of 0.025 for ViT-Base. We train the linear classifier for 90 epochs. The learning rate is linearly warmed up for the first 10 epochs, and decayed to zero by a cosine learning rate schedule. We do not use mixup, cutmix, drop path, or color jittering, and set the weight decay to zero. For other methods, we report the number from original papers.

Model	Epochs	LP Acc.(%)
contrastive method		
MoCov3 [4]	300	76.7
DINO [2]	400	78.2
MAE method		
BEiT [1]	800	56.7
SimMIM [14]	800	56.7
MAE [6]	1600	68.0
CAE [3]	800	68.3
MVP [10]	300	75.4
MILAN [7]	400	78.9
BEiT v2 [9]	300	80.1
FD-CLIP [11]	300	80.3
Ours	200	79.9

Table 1. Comparison of the linear probing top-1 accuracy (LP Acc.) on ImageNet-1K dataset. "Epochs" refer to the pre-training epochs of various methods.

From the results, MaskAlign outperforms methods based on contrastive learning, such as MoCov3 and DINO, and masked modeling methods based on CLIP, such as MVP and MILAN.

B. ADE20K Semantic Segmentation

We transfer our pre-trained backbone models to semantic segmentation task on the ADE20K dataset [15]. Following MAE, the ViT models pre-trained on ImageNet-1K dataset serve as the backbone of UperNet [13], and are finetuned together with the segmentation layers. We report the mean intersection over union (mIoU) averaged over all semantic categories.

Model	Epochs	mIoU(%)
contrastive method		
MoCov3 [4]	300	47.3
DINO [2]	400	47.2
MAE method		
BEiT [1]	800	45.7
MAE [6]	1600	48.1
CAE [3]	1600	50.2
PeCo [5]	300	46.7
MVP [10]	300	52.4
MILAN [7]	400	52.7
Ours	200	52.1

Table 2. Comparison of the semantic segmentation on ADE20K dataset. "Epochs" refer to the pre-training epochs of various methods.

From the results, MaskAlign achieves comparable segmentation performance with MVP, MILAN, but with much fewer pre-training epochs.

C. ImageNet-9 Backgrounds Challenge

We add the ImageNet-9 [12] Backgrounds challenge under linear probing setting. We follow the protocol of AttMask [8] and report results on: Only-FG (OF), Mixed-Same (MS), Mixed-Rand (MR), and Mixed-Next (MN), No-FG (NF), and original.

MaskAlign has better robustness than MILAN. As Background has a higher probability of being masked and reconstructed (due to SAS in MILAN), the removal of reconstruction makes the model focus more on the foreground.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 1

Model	OF	MS	MR	MN	NF	Original
AttMask [8]	75.2	76.2	62.3	59.4	40.6	89.8
MAE [6]	81.3	77.8	66.3	64.0	38.6	91.9
MILAN [7]	**89.2**	87.1	77.9	74.7	46.8	96.2
Ours	87.7	**87.2**	**78.6**	**76.6**	**51.9**	**96.4**

Table 3. Comparison of the linear probing results on ImageNet-9 dataset.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1

[3] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 1

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 1

[5] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 1

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

[7] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 1, 2

[8] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022. 1, 2

[9] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1

[10] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *ECCV*, 2022. 1

[11] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 1

[12] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 1

[13] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 1

[14] 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

[15] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 1