SdAE: Self-distillated Masked Autoencoder Supplementary Material

Yabo Chen[1†], Yuchen Liu[2†], Dongsheng Jiang[3†], Xiaopeng Zhang[3] , Wenrui Dai[1], Hongkai Xiong[2], and Qi Tian[3⋆]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Department of Electronic Engineering, Shanghai Jiao Tong University
{chenyabo, liuyuchen6666, daiwenrui, xionghongkai}@sjtu.edu.cn
[3]Huawei Cloud EI
dongsheng_jiang@outlook.com, zxphistory@gmail.com, tian.qi1@huawei.com

1 Implementation Details

Pretraining. The settings are almost the same as MAE [9]. We use AdamW for optimization and train the CAE for 300 epochs with batch size 768. We set the learning rate as 8e-4, with cosine learning rate decay and a 60 epoch warmup, and set the weight decay as 0.05. We employ the drop path with the ratio of 0.25 only on the encoder. The momentum coefficient is set as 0.96 and with a cosine schedule to 0.99, and EMA is conducted per pre-training epoch. We set the mask ratio as 0.75 with only 49 tokens fed into the student branch. The masked tokens are divided into 3 folds where each fold contains 49 tokens, and all folds are fed into a shared weighted teacher branch.

Fine-tuning on ImageNet. We follow the fine-tuning setting almost the same as MAE [9] to use layer-wise learning rate decay, weight decay, and AdamW. The batch size is 2048, and the weight decay is 0.05. For ViT-B, we train 100 epochs with base learning rate 5e-4, layer-wise decay rate 0.65, drop path rate 0.1, and warmup epoch 10. For ViT-L, we train 50 epochs with base learning rate 1e-3, layer-wise decay rate 0.75, drop path rate 0.2, and warmup epoch 5.

Object Detection and Instance Segmentation on COCO. We utilize the same setting as CAE [5] that uses multi-scale training and resizes the image with the size of the short side between 480 and 800 and the long side no larger than 1333. The batch size is 32, the learning rate is 3e-4, and the layer-wise decay rate is 0.75. We train the network with the $1\times$ schedule: 12 epochs with the learning rate decayed by $10\times$ at epochs 9 and 11. We do not use multi-scale testing. The Mask R-CNN implementation follows MMDetection [4].

Semantic Segmentation on ADE20K. We utilize the same setting as CAE [5]. We use AdamW as the optimizer. The batch size is 16 and the layer-wise decay rate is 0.65. The input resolution is 512×512. We use the learning rates as 4e-4 for all the results in our experiments. We conduct fine-tuning for 160 K steps, and we do not use multi-scale testing.

⋆ Correspondence to Qi Tian. †Equal contribution.

Table S-1. Image classification results on the ILSVRC-2012 ImageNet dataset with top-1 accuracy. "Epochs" refers to the number of pre-training epochs. MoCo v3 adopts multi-crop augmentation with 2 global crops of 224 × 224 for pre-training.

Method	Epochs	Crops	Accuracy
Methods using ViT-L:			
Train from Scratch	300	–	82.6
MoCo v3	300	2	84.1
BEiT	1600	1	85.2
iBOT	250	1	85.0
MAE	400	1	84.3
MAE	1600	1	**85.9**
MaskFeat	1600	1	85.7
SdAE	300	1	85.7

2 More Results for Larger Models and Longer Pre-training Epochs

SdAE can also perform well with only 300 epochs pre-training on a larger model scale such as ViT-L. We study the fine-tuning on the ILSVRC-2012 ImageNet dataset [10] with 1k classes and 1.3M images. For a fair comparison, we directly follow most of the hyperparameters of MAE [9] in our fine-tuning experiments. All reported experimental results are only fine-tuning for 50 epochs.

As shown in Table S-1, compared with the models trained by random initialization (train from scratch), our pre-trained SdAE significantly improves the performance. Specifically, vision transformers trained from scratch only achieve 82.6% top-1 accuracy with ViT-L. While our SdAE achieves 85.7%, demonstrating the effectiveness of pre-training with unlabeled data.

Compared with previous self-supervised methods for vision transformers, our proposed SdAE surpasses them on ImageNet fine-tuning by a large margin. For ViT-L, our SdAE outperforms MoCo v3 by 1.6% top-1 accuracy with the same number of training epochs and our SdAE outperforms MAE by 1.4% top-1 accuracy with the less number of training epochs. Besides, our SdAE outperforms BEiT by 0.5% top-1 accuracy, while BEiT requires an additional pre-trained codebook and longer training epochs. In addition, our SdAE outperforms iBOT by 0.7% top-1 accuracy. Moreover, compared to the 1600 epoch pre-trained ViT-L of MAE and MaskFeat, which requires very large computational costs, our SdAE can achieve comparable performance.

As shown in Table S-2, although the cost of SdAE *de facto* surpasses MAE per epoch, it can speed up convergence and achieve comparable performance in fewer epochs. It is also more efficient than SimMIM that adopts the mask token for the encoder. For longer pre-training epochs, SdAE is faced with a little bit performance degradation on ImageNet fine-tuning. We speculate that this is due to the fact that the Vit-base capacity is relatively close to the performance upper bound of the MIM tasks. However, SdAE shows continuous performance

Table S-2. Longer pre-training epochs results with overall computational costs and memory cost compared with MAE, SimMIM and CAE on ViT-Base using NVIDIA V100 GPUs. Ft: Image classification on the ImageNet dataset with top-1 accuracy under fine-tuning 100 epochs. Det: object detection on the COCO dataset, and AP^b is reported. Seg: semantic segmentation on the ADE20K dataset, and mIoU is reported.

Method	Epoch	Train Time	GPU Mem	Ft	Det	Seg
MAE	300	45.0h	14.93G	82.9	45.4	45.8
CAE	300	-	-	83.3	48.0	47.7
SdAE	300	60.8h	16.02G	**84.1**	**48.9**	**48.6**
SimMIM	800	188.6h	20.32G	83.8	46.5	46.9
MAE	800	140.1h	14.93G	83.4	47.8	47.3
CAE	800	-	-	83.6	49.2	48.8
MAE	1600	278.5h	14.93G	83.6	48.4	48.1
SdAE	800	174.1h	16.02G	**84.0**	**49.7**	**49.0**

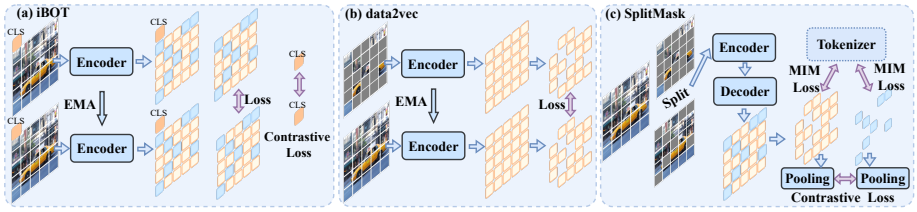

Fig. S-1. Comparison of other recently proposed generative-based self-supervised learning methods. (a) iBOT heavily relies on contrastive loss. (b) data2vec fails to consider the redundancy existing in the mask and visible tokens. (c) SplitMask fails to consider using a two-branch distillation structure and still needs an extra tokenizer.

enhancement with longer training epochs on ADE20K semantic segmentation and COCO object detection, which also surpasses other methods by a considerable margin.

3 Comparison of iBOT, data2vec, and SplitMask

Except for the comparison of typical generative-based self-supervised learning methods in Fig. S-1 such as BeiT [2], PeCo [7], MAE [9] and CAE [5], we also provide the comparison of several recently proposed works in Fig. S-1.

iBOT [11] is more likely a contrastive learning/instance discrimination-based method. iBOT needs careful parameter setting of multi-crop augmentation, which uses 10 local crops with local scale being (0.05,0.32) and global scale being (0.32,1.0). In addition, iBOT heavily depends on the contrastive loss. MIM without the class token contrastive loss leads to undesirable results of 9.5% kNN accuracy and 29.8% linear accuracy on iBOT, indicating that iBOT benefits more from contrastive structure than MIM to extract the visual semantics.

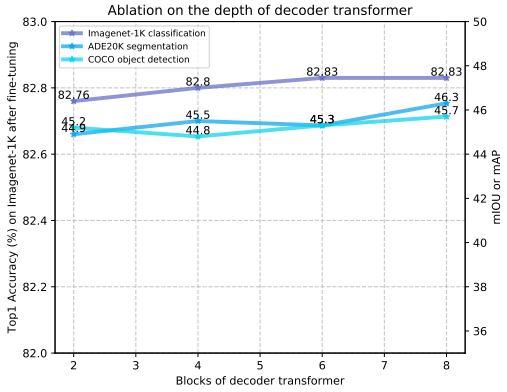

Fig. S-2. Ablation studies on the depth of the decoder transformer.

Data2vec [1] also uses an EMA parameterization of the two-branch teacher-student distillation structure. However, data2vec does not consider the spatial redundancy existing in the network structure. Not only visible unmasked patches but also learned mask embedding tokens are fed into the student branch. Furthermore, the whole input image will be fed into the teacher branch, ignoring the reconstruction loss computed only on masked tokens, which will also increase computational costs. In addition, data2vec conducts an EMA update per iteration. The MIM output and reconstruction target will be very similar if the momentum coefficient is not extremely small. So the network is sharply sensitive to the momentum coefficient. So, data2vec needs precisely tuning on this coefficient where in ViT-L they need to first set momentum coefficient as 0.9998 for the first 800 epochs and then reset the learning rate schedule and the teacher weights to the student and continue for another 800 epochs with momentum coefficient as 0.9999.

SplitMask [8] considers the redundancy of split tokens. However, SplitMask still needs an additional tokenizer to produce discrete latent representations to conduct MIM. Moreover, SplitMask does not consider using a two-branch network to distill the representation between split tokens but adds a pooling module to calculate the contrastive loss between global representations.

3.1 Ablations on the Depth of Decoder

The decoder of the autoencoder, which maps the latent representation back to the reconstruction space, plays an essential role in the masked image modeling task. In the language MLM, the decoder predicts missing words that contain rich semantic information so that the decoder can be trivial (an MLP) in BERT [6]. However, in MAE [9], the decoder reconstructs the image pixels, which reconstructs the latent representations into the low-level pixel space. Thus, MAE

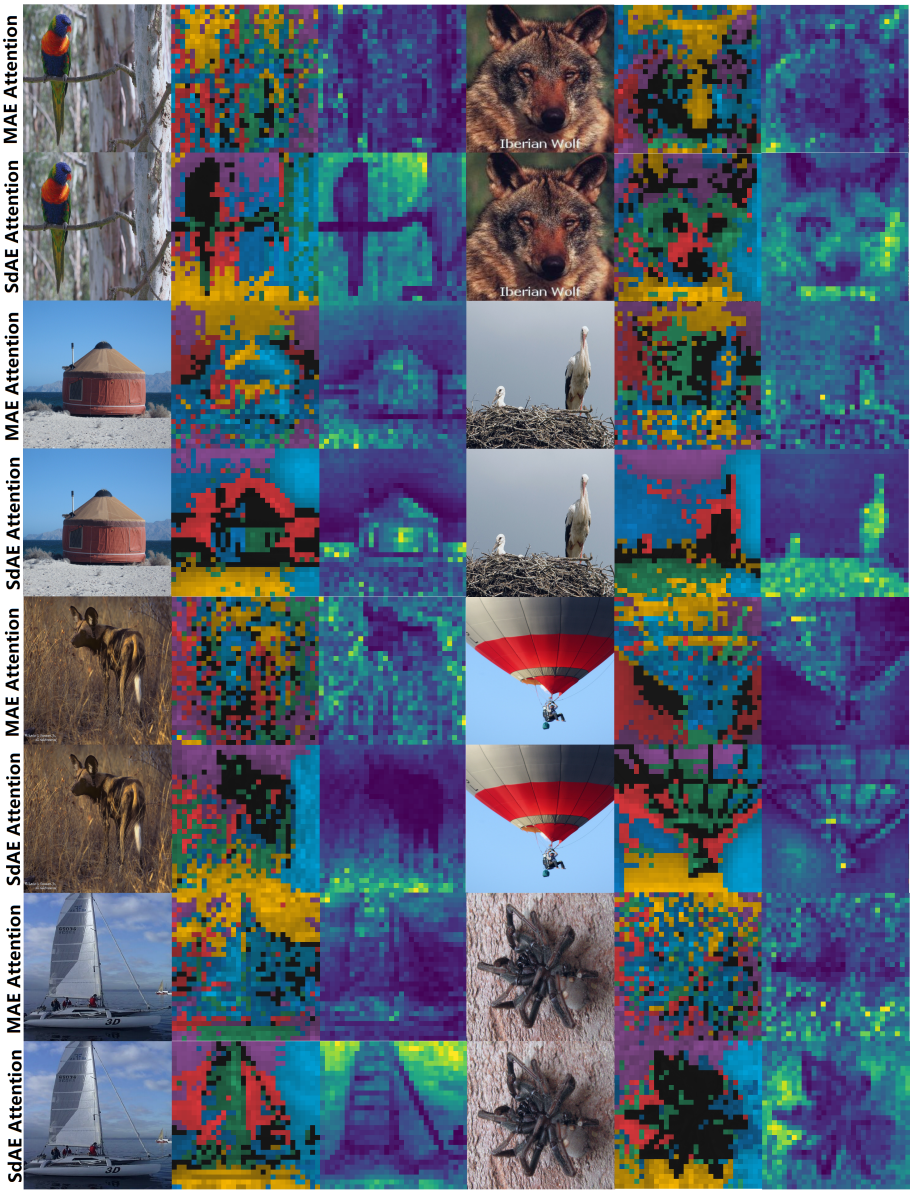

Fig. S-3. Attention maps visualization of MAE and SdAE. The first column and forth column place the original images. The second column and fifth column are the visualization of different heads from the last layer with different colors. The third column and sixth column are the visualization of mean of all attention heads.

requires a relatively powerful decoder. In comparison, our student network maps the latent representations to the high-level semantic features so that the decoder

can be lighter. That is another potential advantage of SdAE. As shown in Fig. S-2, the experiment shows that the depth of the decoder has little impact on the performance. Specifically, even with two layers of the decoder transformer, our SdAE achieves 82.76% top-1 accuracy and 45.2 mAP on COCO object detection, which only suffers 0.07% and 0.5 mAP performance degradation compared with eight layers of the decoder transformer.

4 Visualization

To analyze, we visualize the self-attention map with 300-epoch pre-trained ViT-B/16 of both MAE and SdAE. We choose the class token as the query and visualize attention maps from different heads of the last layer with different colors, following iBOT [11]. As shown in Fig. S-3, we indicate that SdAE shows the capability to learn high-level semantic features to separate different parts of objects. Compared with MAE, SdAE is able to learn more meaningful high semantic information.

Specifically, in the figure, we observe SdAE can distinguish the bird from the tree or distinguish the eyes and ears of the Iberian wolf. Moreover, SdAE can also focus on the discriminative details of the object (e.g., the skeleton of a hot air balloon and sailboat rope) without using the contrastive loss. For more complex scenes like spiders on the surface of complex texture SdAE is still able to distinguish subjects. This is because SdAE does not need to reconstruct every pixel, so it did not pay attention to useless details.

With only a simple normalized feature MSE loss, we can achieve similar behaviors with intricately designed instance discrimination methods such as DINO [3].

References

1. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022)
2. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv: Computer Vision and Pattern Recognition (2021)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning (2022)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2018)

7. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Peco: Perceptual codebook for bert pre-training of vision transformers (2022)
8. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? arXiv e-prints (2021)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
11. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer (2022)