Generic-to-Specific Distillation of Masked Autoencoders
Supplemental Material

Wei Huang[1,*], Zhiliang Peng[1,§,*], Li Dong[2], Furu Wei[2], Jianbin Jiao[1,†], Qixiang Ye[1,†]
University of Chinese Academy of Sciences[1]
Microsoft Research[2]

In this supplementary material, more experimental details are provided. Please refer to the ZIP file for the code of this study.

1. Hyperparameters

1.1. Image Classification

For distillation, as in [7], we added a learnable distillation token, which is combined with the cls token to produce final predictions in the inference phase. In experiments, the data augmentation and optimizer follow the fine-tuning recipe of MAE [3], while the learning rate, training epochs and layer-wise learning-rate decay are specified. For models training from scratch (e.g., DeiT⚗), we set the layer decay value as 1.0, which means no layer decay is adopted. For pre-trained models (e.g., MAE [3], G2SD), we set the layer decay value to 0.75 and training epochs to 200.

Table 1. Hyperparameters for distilling on ImageNet-1K.

Hyperparameters	**Value (Fine-tuning)**	**Value (From scratch)**
Training epochs	200	500
Base learning rate	1e-3	2.5e-4
Layer decay	0.75	1.0
Warm up epochs	5	
Label smoothing	0.1	
Mixup	0.8	
Cutmix	1.0	
Drop path	0.0	
Batch size	1024	
Weight decay	0.05	
Optimizer	AdamW	
Learning rate schedule	Cosine decay	
Augmentation	RandAug(0,0.5)	
Optimizer momentum	β_1, β_2 = 0.9, 0.999	

1.2. Object Detection and Instance Segmentation

In the experiments, we adopt the official codebase[1] and follow the settings used in ViTDet [6]. The total batch size is set to 64 (8 images per GPU). The learning rate is set to $1e^{-4}$, the backbone's drop path rate is 0.1, and the distill warm step is 500. The overall training target is the same as [2]: $L = L_{GT} + \alpha L_{FPN} + \beta L_{head}$, where α and β are respectivvely set to 0.001 and 0.1.

1.3. Semantic Segmentation

In this experiment, we adopt the BEiT's segmentation codebase[2] and set the total batch size to 32 (4 images per GPU). The backbone's drop path rate is 0.1. The layer decay rate is 0.75. The learning rate of ViT-Small and ViT-Tiny are respectively set to $2e^{-4}$ and $5e^{-4}$. We set the temperature parameter $\tau = 1$, the loss weight $\alpha = 3$ for the logits map distillation.

2. Training Time and Efficiency

As shown in Table 2, G2SD outperforms DeiT [7] and DeiT⚗ [7], which have a longer training schedule (500 epochs). The teacher of DeiT⚗ is the same as G2SD's. In the generic distillation stage, since the input of G2SD is a masked image (75% patches are discarded), the training time per epoch is less than DeiT (which computes the whole image).

Table 2. G2SD vs DeiT. The total training epochs is 500.

Methods	**1-st stage**	**2-nd stage**	**Time**	**Top-1 Acc (%)**
G2SD	G.D 300 epochs	S.D 200 epochs	71 h	82.5
DeiT⚗	Supervised+Distillation 500 epochs		112 h	81.7 (**-0.8**)
DeiT	Supervised 500 epochs		53 h	81.4 (**-1.1**)

[*] Equal contribution. § Contribution during internship at Microsoft Research. † Corresponding authors.

[1] https://github.com/facebookresearch/detectron2/tree/main/projects/ViTDet

[2] https://github.com/microsoft/unilm/beit

3. Detection Performance with ViTDet

For the lack of official Mask-RCNN [4] results and checkpoints of MAE [3], we choose ViTDet [6] as the detector. In Table 3, the backbone models are initialized from various supervisions, *e.g.*, supervised methods (DeiT [7]), distilled methods (DeiT⚗ [7] and G2SD) and self-supervised methods (DINO [1] and iBoT [8]). From Table 3, G2SD significantly outperforms competitors on performance and convergence speed.

Table 3. Performance on MS COCO using the ViTDet framework [6], which is trained for 100 epochs with single-scale input (1024×1024).

Methods (Supervision)	ImageNet Acc (%)	AP^{bbox}	AP^{mask}
DeiT-S (sup., 300e)	79.9	45.7	40.7
DeiT-S⚗ (sup.&distill., 300e)	81.2	47.2	41.9
DeiT-S (sup., 500e)	81.4	46.9	41.6
DINO-S (self-sup., 3200e)	82.0	49.1	43.3
iBOT-S (self-sup., 3200e)	82.3	49.7	44.0
G2SD-S (w/o S.D, 300e)	82.0	49.9	44.5
G2SD-S (300e)	82.5	50.6	44.8

4. More Ablations

Target Configuration. In the paper, we conducted ablation studies on intermediate features as generic distillation targets. Compared with using intermediate features as distillation targets, taking the teacher's prediction as distillation objective [5, 7] is also a popular alternative. Therefore, we take the MAE's predictions as the generic distillation targets in Table 4. When taking the MAE's predictions as the targets for masked positions, the performance drops to 81.4% (without specific distillation) and 81.8% (with specific distillation). This observation is consist with the results in Table 5 (*bottom*), where the last several layers in decoder are more specialized for low-level information reconstruction task.

Mask Ratio. A high mask ratio (75%) works well in MAE [3], but the suitable mask ratio in generic distillation still needs to be explored. In general, predicting masked features is more challenging than predicting pixels. However, the observations are consistent with the teacher MAE, as illustrated in Tab. 5 (*top*), where a high mask ratio tends to generate good results. The reason may be that the teacher model can express itself to the greatest extent when the mask ratio is consistent with the MAE pre-training phase.

Table 4. Ablation study of distillation targets on ImageNet-1k. 'S.D' is short for specific distillation.

Distillation targets	W/O S.D Acc (%)	W S.D Acc (%)
Our default settings	**82.0**	**82.5**
MAE's reconstructions	81.4	81.8
MAE's reconstructions + GT	81.5	81.7

Table 5. Ablation on the mask ratio.

Mask ratio	0.05	0.25	0.55	0.75	0.9
Top-1 Acc(%)	81.7	81.7	81.6	**82.0**	81.8

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 2

[2] Zhixing Du, Rui Zhang, Ming-Fang Chang, Xishan Zhang, Shaoli Liu, Tianshi Chen, and Yunji Chen. Distilling object detectors with feature richness. In *NeurIPS*, 2021. 1

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

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE ICCV*, 2017. 2

[5] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[6] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ArXiv*, abs/2203.16527, 2022. 1, 2

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877*, 2020. 1, 2

[8] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2