Supplementary Materials: Attention Distillation: self-supervised vision transformer students need more guidance

Kai Wang
kwang@cvc.uab.es
Fei Yang (*corresponding author)
fyang@cvc.uab.es
Joost van de Weijer
joost@cvc.uab.es

Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain

1 Table 2 visualization

A graphical visualization for Table 2 in the main paper is shown in Fig. 6, where you can easily observe the performance difference between each teacher and student pair. We can also observe that the performance drop from linear probing to k-NN is effectively reduced by AttnDistill.

2 Networks configurations

In this paper, our networks configurations mainly refer to the ViT designs in DINO [1], DEiT [9], Mugs [12] and iBOT [11], where the number of heads H and position embedding (PE) strategy (learnable PE and ViT-S with 6 head) are different from MoCo v3 [3] (fixed sin-cos PE and ViT-S with 12 head). The detailed networks configurations are shown in Table 7.

3 Additional Implementations

3.1 Pre-Training recipe

A clear pre-training recipe is shown in Table 8. We mainly refer to the training recipe from MAE [5].

3.2 More details for evaluations

***k*-NN, Linear Probing and finetuning on ImageNet-1K.** To evaluate the quality of pre-trained features, we either use a k-nearest neighbor (k-NN) classifier or a linear classifier

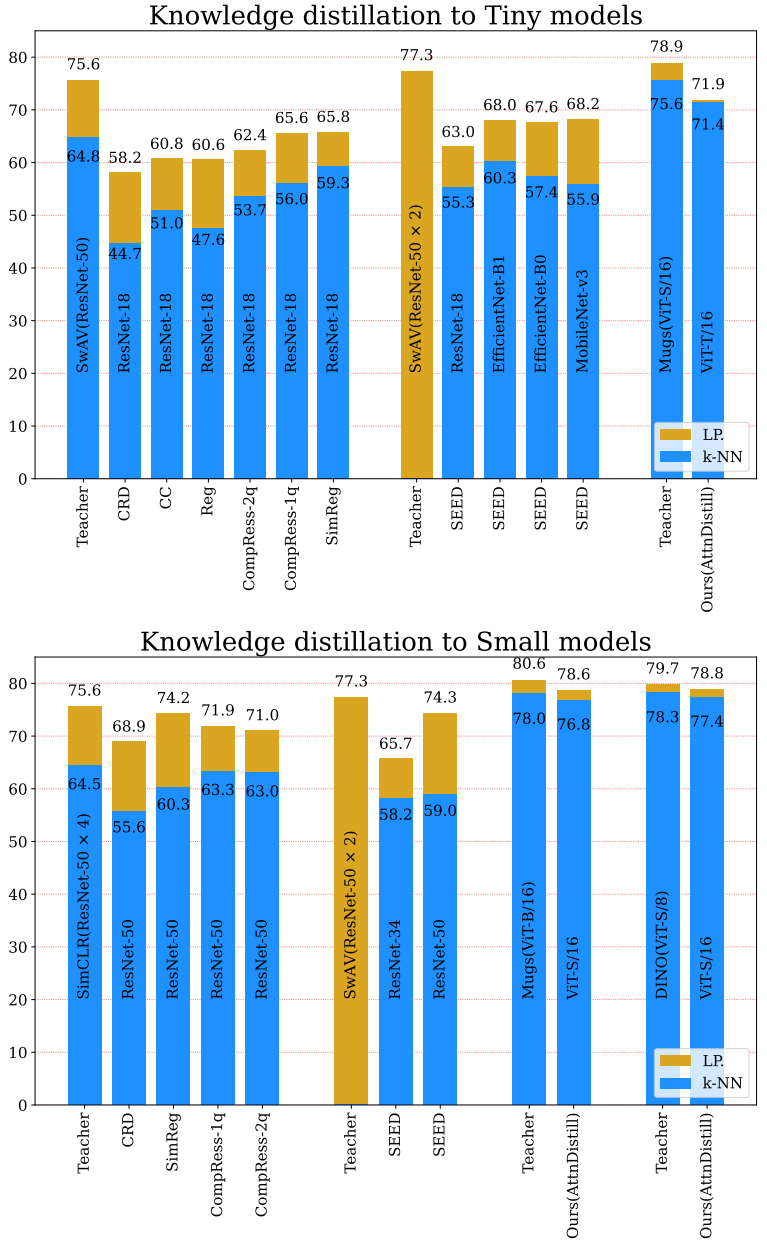

Figure 6: Visualization of Table 2 results, providing both the linear probing and k-NN accuracies for multiple methods. Note that AttnDistill significantly reduces the gap with the teacher model.

Networks configurations for experiments on **ImageNet-1K**								
	PE	model	layers	dim	heads	patch size	#tokens	#params
Teacher	learnable	Mugs (ViT-S/16)	12	384	6	16	197	22M
		DINO (ViT-S/8)	12	384	6	8	785	22M
		Mugs (ViT-B/16)	12	768	12	16	197	85M
Student	learnable	AttnDistill (ViT-T/16)	12	192	3	16	197	5.7M
		AttnDistill (ViT-S/16)	12	384	6	16	197	22M
Networks configurations for experiments on **ImageNet-Subset**								
Teacher	sin-cos	MAE (ViT-S/16)	12	384	6	16	197	22M
Student	sin-cos	AttnDistill (ViT-T)	12	192	6	16	197	5.7M
		AttnDistill (ViT-T)	12	192	3	16	197	5.7M
		AttnDistill (ViT-T)	12	192	3	32	65	5.7M
		AttnDistill (ViT-T)	8	192	3	16	197	3.8M
		AttnDistill (ViT-T)	8	192	3	32	65	3.8M

Table 7: **Networks configuration.** "layers" is the number of Transformer blocks, "dim" is channel dimension and "heads" is the number of heads in multi-head attention. "# tokens" is the length of the token sequence, "# params" is the total number of parameters. "PE" is the position embedding strategy. We consider 224×224 resolution inputs.

config	value
optimizer	AdamW [8]
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$ [2]
batch size	4096
learning rate schedule	cosine decay [7]
warmup epochs	40
training epochs	500 (ViT-T/16 ImageNet-1K) 800 (ViT-S/16 ImageNet-1K) 3200 (ViT-T/16 ImageNet-Subset)
augmentation	RandomResizedCrop

Table 8: Pre-Training settings for ViTs distillation on **ImageNet-1K** and **ImageNet-Subset**.

config	value
optimizer	LARS [10]
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 9: **Linear probing setting** on **ImageNet-Subset** for self-supervised knowledge distillation with a MAE(ViT-S/16) teacher.

on the frozen representation. We follow the evaluation protocols in DINO [1], iBOT [11], Mugs [12]. For k-NN evaluation, we sweep over different numbers of nearest neighbors. For linear probing and finetuning evaluations, we sweep over different learning rates.

Semi-supervised learning on ImageNet-1K. In this setting, first, models are trained self-supervised on all *ImageNet-1K* data. Next, labels for a small fraction of data (1% or 10%) are used to perform fine-tuning, linear probing or *k*-NN classification. This is also an *extension* to the semi-supervised learning evaluations in DINO, iBOT and Mugs papers. For k-NN evaluation, we sweep over different numbers of nearest neighbors. For linear evaluation, we sweep over different learning rates. For fine-tuning evaluation, we fine-tune the pre-trained backbone for 1000 epochs with learning rate set to 5e-6.

Transfer learning. We pretrain the model on ImageNet-1K, and then fine-tune the pre-trained backbone on various datasets with the same protocols and optimization settings as in DINO, iBOT, and Mugs. For both ViT-T and ViT-S, we use AdamW optimizer with a minibatch of 1024, we fine-tune the pretrained model 1000 epochs by sweeping the learning rates. The weight decay is fixed to be 0.05.

Linear probing evaluation on ImageNet-Subset Since on ImageNet-Subset we consider the MAE as the teacher, thus in linear probing, we also follow the evaluation protocols from MAE [5] as shown in Table 9.

4 MAE pretraining on ImageNet-Subset

The linear probing curves of MAE [5] pre-training on ImageNet-Subset is shown in Fig. 7 for further references (from 1200 epochs to 3200 epochs). We can observe that both ViT-S(12-layers, 6-heads, 16-patches, 384-dim) and ViT-T(12-layers, 6-heads, 16-patches, 192-dim) both are saturated at 3200 epochs.

5 Additional ablation studies.

To prove the generalizability of AttnDistill, we perform the ablation study on ImageNet-Subset with a fixed MAE(ViT-S/16) teacher and vary the architecture of the student model. Apart from the ablation studies shown in Fig.5 in the main paper, here we extend the ablation studies and also display all the numbers included in Fig.5. As can be seen from Table 10, our ablation study can be roughly divided into six parts. The beginning three parts (a)-(c) are corresponding to the Fig.5(a)-(c). Except that, we further show our ablation study on the following three aspects:

- The design of the linear mapping $\mathcal{P}$: In Table 10-(d), we vary the number of $\mathcal{P}$ layers in {1,2,4,8} and evaluate the output features from each layer. As can be seen that, the output feature before $\mathcal{P}$ (indicated by a 0 in column *Evaluation P layer*) is always a good choice for all considered layer-numbers variations. Also, for the number of layers 4 is a better choice.
- The hyperparameter of AttnDistill: In Table 10-(e) and Fig. 8, we ablate the λ and T. The optimal choice is $\lambda = 0.1, T = 10.0$.

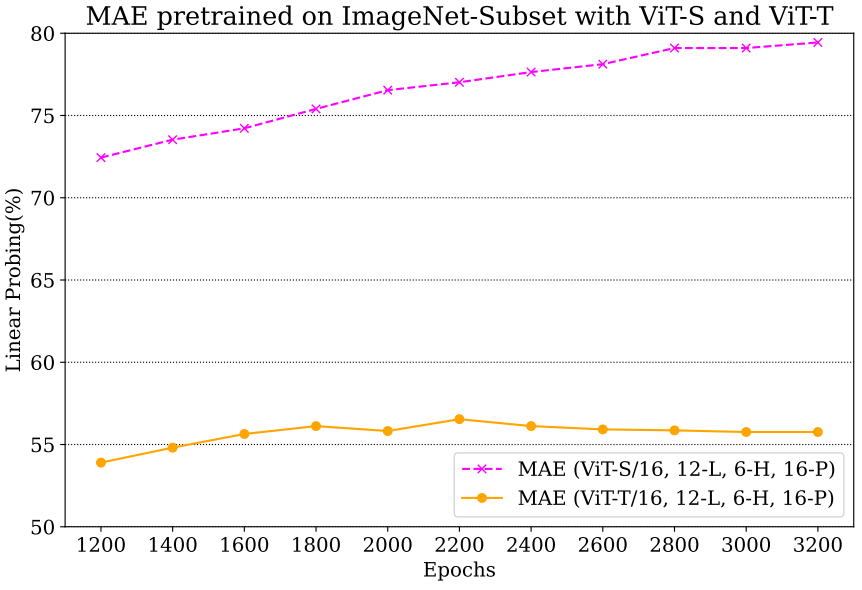

Figure 7: MAE pretrained on ImageNet-Subset with ViT-S and ViT-T models.

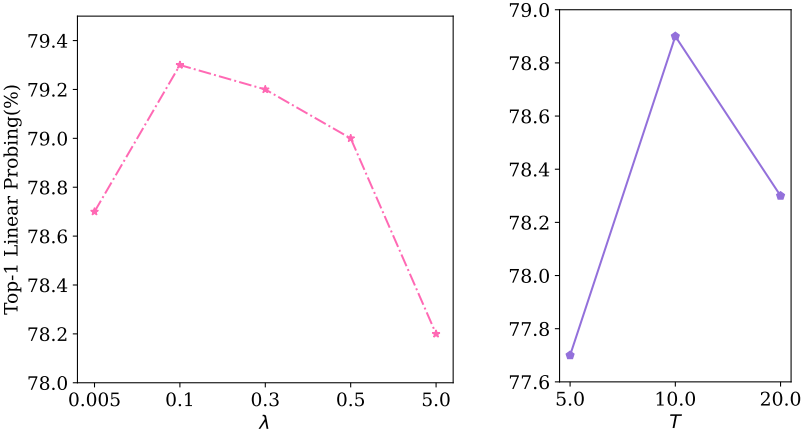

Figure 8: Ablation study over T and λ.

- Other ablation studies: here we fix the design of the student model and vary the strategy to compute the attention distillation loss. We imitate the ATS [4] to compute the token scores with the norm of *Value* vectors in the MSA module of ViTs. We also replace the KL divergence in $\mathcal{L}_a$ with MSE loss. Neither of them could work better than our solution.

	Study on	Layers	Heads	Patch size	**PA**	Patch token align	CompRess KD	$\mathcal{P}$ layers	Attn Layer	**AttnDistill**	Attn Aggr	Attn Interpolate	Attn λ	Attn T	Evaluation $\mathcal{P}$ layer	Top-1 LP.	Top-5 LP.
Dataset: ImageNet-Subset																	
Teacher: *ViT-S* (12-Layer, 6-Head, 16-Patch, 3200 epo., 384 dim); Top-1 LP.: 79.4; Top-5 LP.: 93.6																	
Student: *ViT-T* (12-Layer, 6-Head, 16-Patch, 3200 epo., 192 dim); Top-1 LP.: 63.7; Top-5 LP.: 86.2																	
(a)	Heads	12	3	16	✓	✗	✗	4	-	✗	✗	✗	-	-	0	77.8	93.6
		12	3	16	✓	✗	✗	4	last	✓	✓	✗	0.1	10.0	0	78.9	93.8
		12	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	0	77.8	93.3
		12	6	16	✓	✗	✗	4	last	✓	✓	✗	0.1	10.0	0	78.9	93.9
		12	6	16	✓	✗	✗	4	last	✓	✗	✗	0.1	-	0	**79.3**	**94.1**
	Patch size	12	6	28	✓	✗	✗	4	-	✗	✗	✗	-	-	0	73.3	91.1
		12	6	28	✓	✗	✗	4	last	✓	✗	✓	0.1	-	0	77.3	93.1
		12	6	32	✓	✗	✗	4	-	✗	✗	✗	-	-	0	73.1	91.0
		12	6	32	✓	✗	✗	4	last	✓	✗	✓	0.1	-	0	77.2	92.8
	Layers	8	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	0	77.7	92.8
		8	6	16	✓	✗	✗	4	last	✓	✗	✗	0.1	-	0	79.1	94.0
		8	3	32	✓	✗	✗	4	-	✗	✗	✗	-	-	0	68.6	88.8
		8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	73.8	91.7
(b)	PA	8	3	32	✓	✗	✗	4	-	✗	✗	✗	-	-	0	68.6	88.8
	AttnDistill	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	73.8	91.7
	MAX	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	73.5	91.6
	MEAN	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	70.6	90.0
	MIN	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	71.9	90.7
(c)	KD	12	6	16	✓	✗	✓	4	-	✗	✗	✗	-	-	0	75.0	92.5
		12	6	16	✗	✗	✓	4	-	✗	✗	✗	-	-	0	71.7	90.7
	Patch	12	6	16	✗	✓	✗	4	-	✗	✗	✗	-	-	0	73.6	91.4
		12	6	16	✓	✓	✗	4	-	✗	✗	✗	-	-	0	76.9	93.4
	Attn Layer	12	6	16	✓	✗	✗	4	all	✓	✗	✗	0.1	-	0	78.7	93.7
		12	6	16	✓	✗	✗	4	last	✓	✗	✗	0.1	-	0	**79.3**	**94.1**
(d)	MLP layers and Eval layer	12	6	16	✓	✗	✗	8	-	✗	✗	✗	-	-	0	76.2	92.4
		12	6	16	✓	✗	✗	1	-	✗	✗	✗	-	-	0	76.3	92.5
		12	6	16	✓	✗	✗	1	-	✗	✗	✗	-	-	1	76.2	92.4
		12	6	16	✓	✗	✗	2	-	✗	✗	✗	-	-	0	76.4	93.2
		12	6	16	✓	✗	✗	2	-	✗	✗	✗	-	-	1	76.4	93.1
		12	6	16	✓	✗	✗	2	-	✗	✗	✗	-	-	2	76.5	93.1
		12	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	0	*77.8*	*93.3*
		12	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	1	77.8	93.3
		12	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	2	77.2	93.2
		12	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	3	76.6	92.9
		12	6	16	✓	✗	✗	4	-	✗	✗	✗	-	-	4	76.2	92.8
(e)	Attn λ	12	6	16	✓	✗	✗	4	last	✓	✗	✗	0.005	-	0	78.7	93.7
		12	6	16	✓	✗	✗	4	last	✓	✗	✗	0.1	-	0	**79.3**	**94.1**
		12	6	16	✓	✗	✗	4	last	✓	✗	✗	0.3	-	0	79.2	94.0
		12	6	16	✓	✗	✗	4	last	✓	✗	✗	0.5	-	0	79.0	93.8
		12	6	16	✓	✗	✗	4	last	✓	✗	✗	5.0	-	0	78.2	93.6
	Attn T	12	3	16	✓	✗	✗	4	last	✓	✓	✗	0.1	10.0	0	78.9	93.8
		12	3	16	✓	✗	✗	4	last	✓	✓	✗	0.1	5.0	0	77.7	93.6
		12	3	16	✓	✗	✗	4	last	✓	✓	✗	0.1	20.0	0	78.3	93.7
(f)	PA	8	3	32	✓	✗	✗	4	-	✗	✗	✗	-	-	0	68.6	88.8
	Weight by |VALUE| [0]	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	70.6	90.5
	MSE loss [0]	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	72.9	90.9
	AttnDistill	8	3	32	✓	✗	✗	4	last	✓	✓	✓	0.1	10.0	0	73.8	91.7

Table 10: Full table for our ablation studies.

6 More visualization of attention maps

Except the Fig.4 in the main paper, here in Fig. 9 and Fig. 10 we also show more visualization of the attention maps obtained from various knowledge distillation methods (MAE-ViT-S →ViT-T) on ImageNet-Subset. More attention visualizations with nearly 1000 images (≈ 10 images per class) are in our supplementary file named *"attn_vis.zip"* with the same layouts.

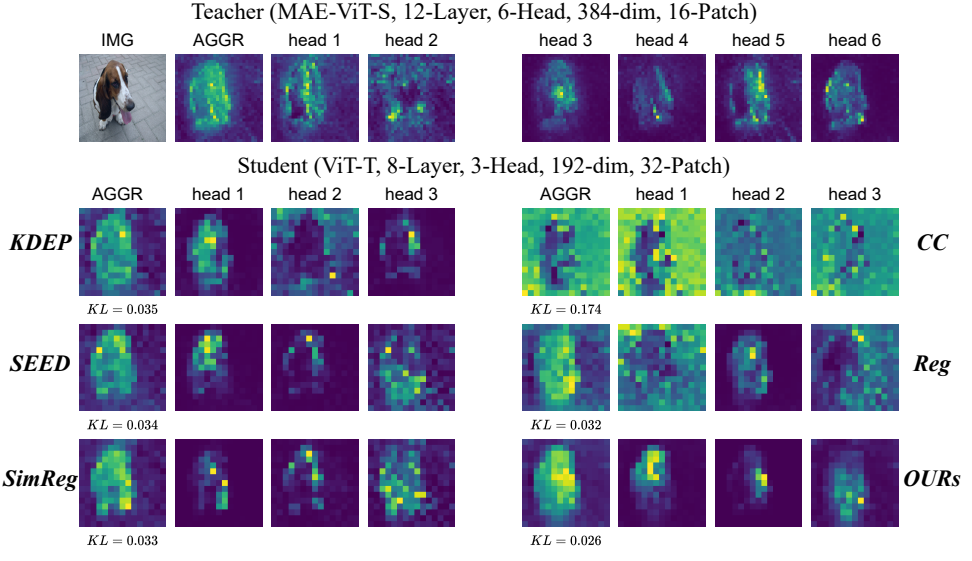

Figure 9: Comparison on attention maps for *"n02088238_194.jpg"*. For the teacher model ViT-S trained with MAE, we show the original image (IMG), the aggregated attention map (AGGR) with our AttnDistill and the attention maps for each head. For the student model ViT-T distilled from the teacher model, we show the aggregated attention map and each head attention map for each method. The KL distances to the teacher aggregated attention maps are shown under each method.

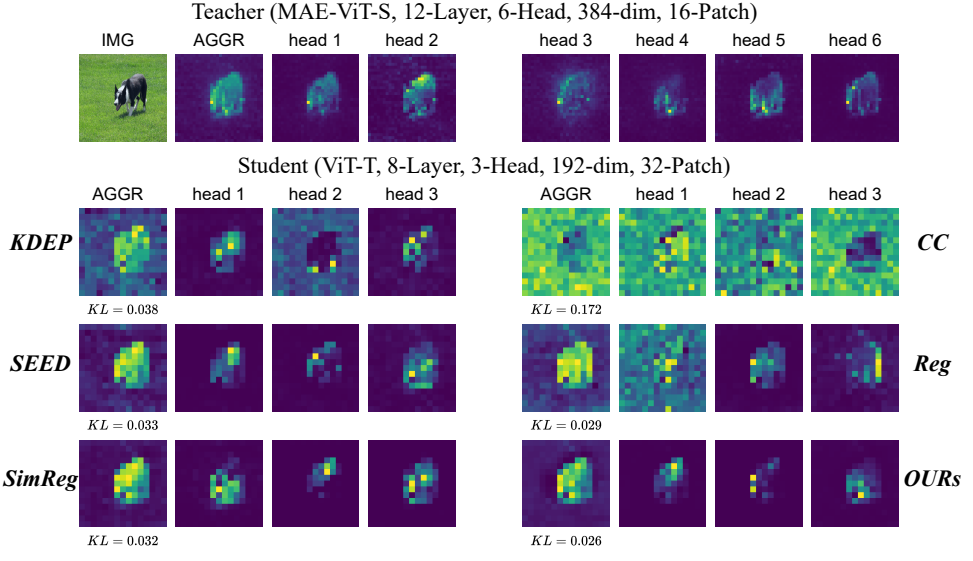

Figure 10: Comparison on attention maps for *"n02106166_45.jpg"*. For the teacher model ViT-S trained with MAE, we show the original image (IMG), the aggregated attention map (AGGR) with our AttnDistill and the attention maps for each head. For the student model ViT-T distilled from the teacher model, we show the aggregated attention map and each head attention map for each method. The KL distances to the teacher aggregated attention maps are shown under each method.

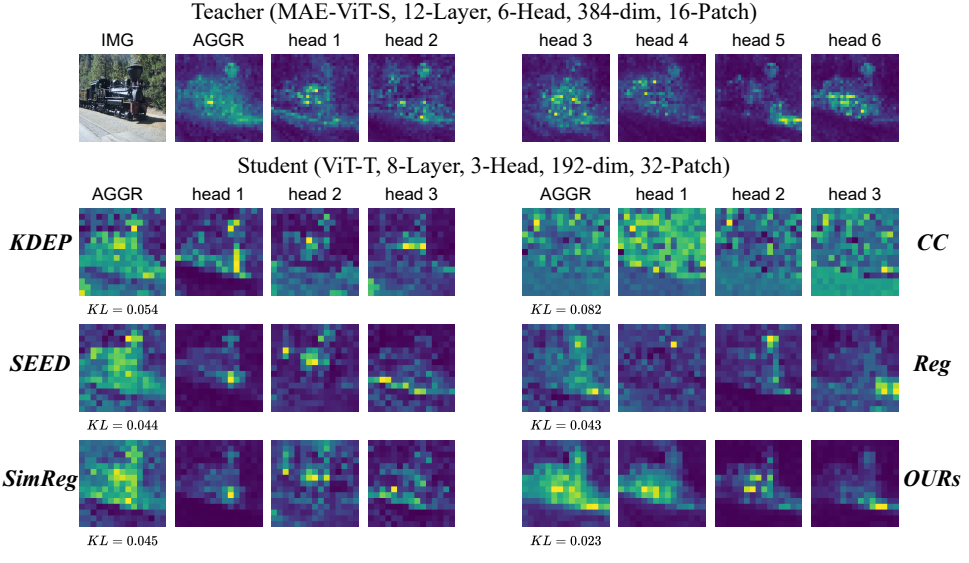

Figure 11: Comparison on attention maps *"n04310018_78.jpg"*. For the teacher model ViT-S trained with MAE, we show the original image (IMG), the aggregated attention map (AGGR) with our AttnDistill and the attention maps for each head. For the student model ViT-T distilled from the teacher model, we show the aggregated attention map and each head attention map for each method. The KL distances to the teacher aggregated attention maps are shown under each method.

References

[1] 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*, pages 9650–9660, 2021.

[2] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[4] Mohsen Fayyaz, Soroush Abbasi Kouhpayegani, Farnoush Rezaei Jafari, Eric Sommerlade, Hamid Reza Vaezi Joze, Hamed Pirsiavash, and Juergen Gall. Ats: Adaptive token sampling for efficient vision transformers. *arXiv preprint arXiv:2111.15667*, 2021.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[6] Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021.

[7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.

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

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021.

[10] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[11] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

[12] Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022.