–Technical Appendix–

DisCo: Remedying Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning

1 Semi-supervised Linear Evaluation

As shown in Fig 1, the semi-supervised linear evaluation results on MobileNet-v3-large is consistent with those on the other small models.

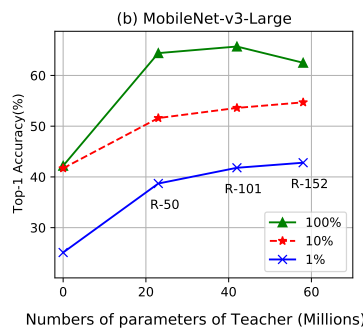

Fig. 1. ImageNet top-1 accuracy (%) of semi-supervised linear evaluation.

2 Transfer to Cifar10

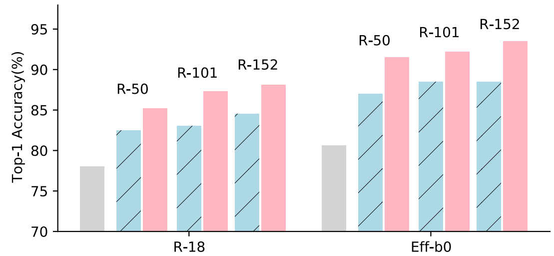

Fig. 2. Top-1 accuracy of students transferred to Cifar10.

Fig 2 shows that DisCo still outperforms the SOTA with a large margin, showing the generalization of learned representation.

3 SwAV as testbed

Table 1. Linear evaluation top-1 accuracy (%) on ImageNet with SwAV as the testbed.

Method	**Eff-b0**	**Mob-v3**
SwAV	46.8	19.4
SwAV + DisCo	62.4	55.7

In order to demonstrate the versatility of DisCo, we further experiment with SwAV as the testbed and teacher is backboned by ResNet-50. The results are shown in Table 1, it can be seen that for models with very few parameters, EfficientNet-B0 and MobileNet-v3-Large, the pre-training results with SwAV are also very poor. When DisCo is utilized, the efficacy is significantly improved.

4 Teacher with Different Pre-training Methods

In order to verify that our method is not picky about the pre-training approach that the teacher adopted, we use three ResNet-50 networks pre-trained with different SSL methods as the teacher under the testbed of MoCo-V2. It can be observed from Table 2 that when using different pre-trained ResNet-50 as teachers, DisCo can significantly boost all the results of small models. Furthermore, with the improvement of the teachers using different and stronger pre-training methods, the results of the student can be further improved.

Table 2. Linear evaluation top-1 accuracy (%) on ImageNet with variants of teacher pre-training methods. All the teachers are ResNet-50 and the first row is student trained by MoCo-V2 directly without distillation, which is baseline.

Teacher		**Student**			
Method	**Acc**	**Eff-b0**	**Eff-b1**	**Mob-v3**	**R-18**
-	-	46.8	48.4	36.2	52.2
MoCo-V2	67.4	66.5	66.6	64.4	60.6
SeLa-V2 71.8	62.2	68.2	66.2	64.1	
SwAV	75.3	70.0	72.1	65.0	65.1

5 Visualization Analysis

In Figure 3, we visualize the learned representations of EfficientNet-B0/ResNet-50 pretrained by MoCo-V2, and EfficientNet-B0 distilled by ResNet-50 using DisCo. For clarity, we randomly select 10 classes from the ImageNet test set and map the learned representations to two-dimensional space by t-SNE [3]. It can be observed that ResNet-50 forms more separated clusters than EfficientNet-B0 when using MoCo-V2 alone, and after using ResNet-50 to teach EfficientNet-B0 with DisCo, EfficientNet-B0 performs very much like the teacher.

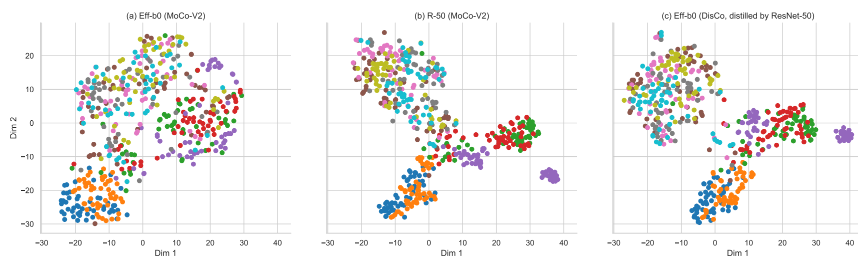

Fig. 3. Clustering results on the ImageNet test set. Different colors represent different classes.

6 More SSL Methods

Table 3. Linear evaluation top-1 accuracy (%) on ImageNet with DINO as testbed. ViT-small[1] and XCiT-small[2] are pre-trained by DINO for 100 epochs.

Teacher Model	Acc	ViT-tiny	XCiT-tiny
-	-	63.2	67.0
ViT-small	77	68.4(5.2↑)	-
XCiT-small	77.8	-	71.1(4.1↑)

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

2. El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers (2021)
3. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. vol. 9 (2008)