Supervised Masked Knowledge Distillation for Few-Shot Transformers

Authors
Han Lin*, Guangxing Han*, Jiawei Ma, Shiyuan Huang, Xudong Lin, Shih-Fu Chang

Observations

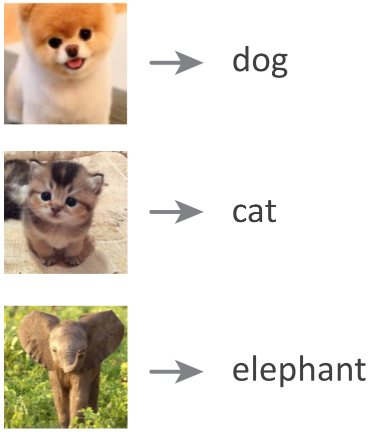
- Unlike CNN-based models, Transformer architecture lacks inductive bias:
  ✔ captures long-range token dependencies
  ✘ data-hungry, easy to overfit to pre-trained datasets which are not large enough

- Overfitting to datasets with insufficiently training data may **hurt the generalization ability** of new classes (e.g., few-shot learning)

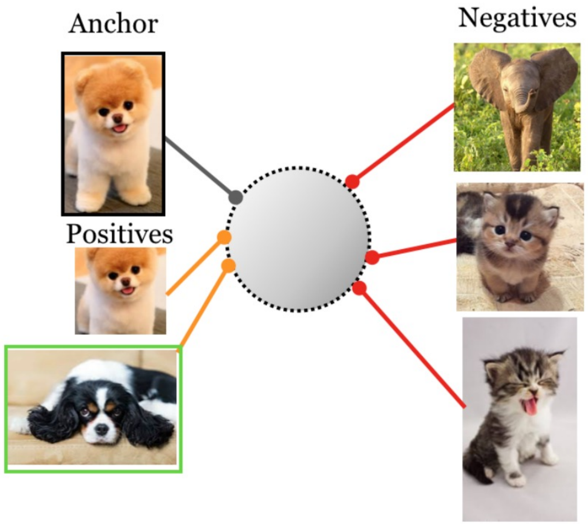
- Our problem settings:
  Transformers on small dataset under few-shot learning

Related Works

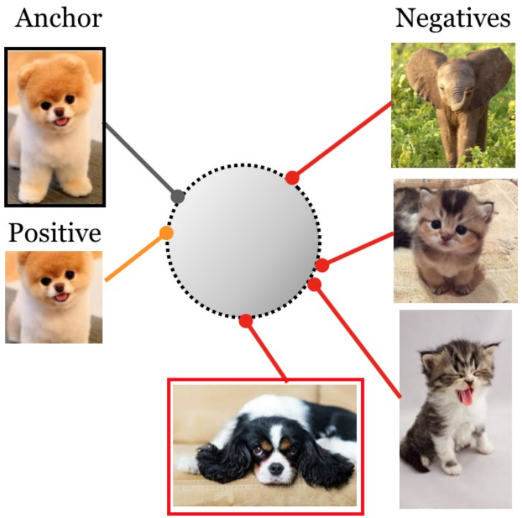
Supervised Learning



Supervised-Contrastive Learning



Self-Supervised Learning



- ✘ Easy to overfit
- ✘ Bad generalization ability

- ✘ Construct negative samples needs large batch size

- ✔ Better generalization ability
- ✔ Can avoid negative samples
- ✘ Hard to capture high-level semantic features

Image source: modified from SupCon [1]

Related Works

- Masked Image Modeling (MIM)
  - Recovering masked pixels from a corrupted input image
  - Combined with self-distillation achieves better performance [3]

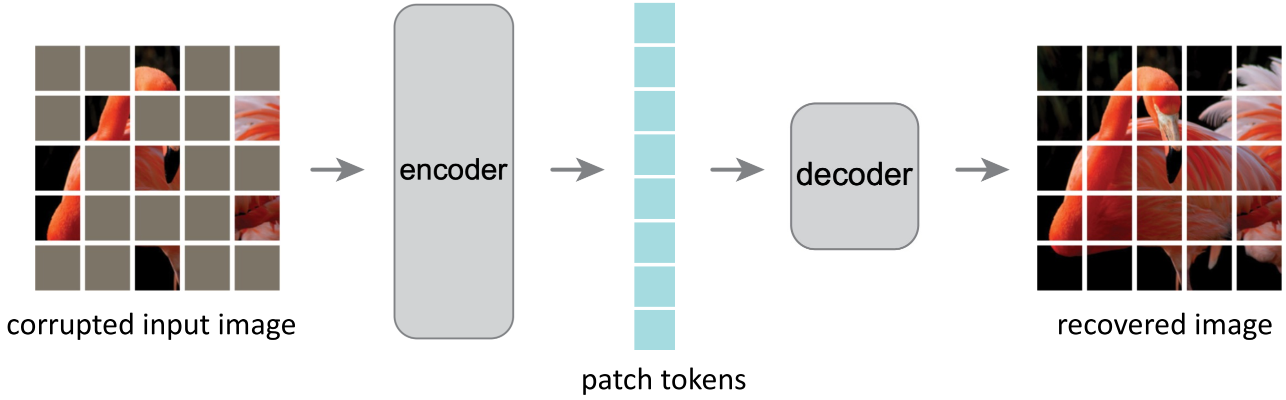


Image source: modified from MAE [2]

Our Work

To tackle the problem of Transformers on small dataset:

class label information ➕ self-supervised knowledge distillation

masked image modeling → supervised masked knowledge distillation framework (SMKD)

Our Training & Few-Shot Evaluation Pipelines:

- Train on base classes:
    - Our method defined in the next page

- Few-shot evaluation on novel classes:
    - Prototype evaluation method
    - N-way K-shot (e.g. 3-way 5-shot)

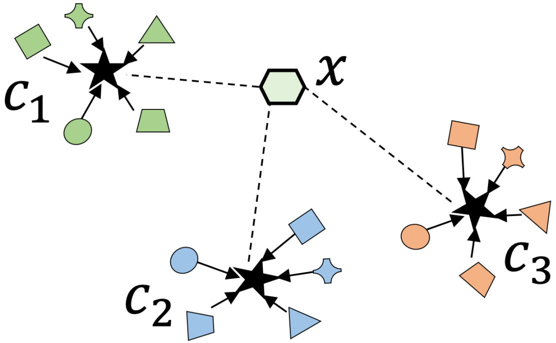
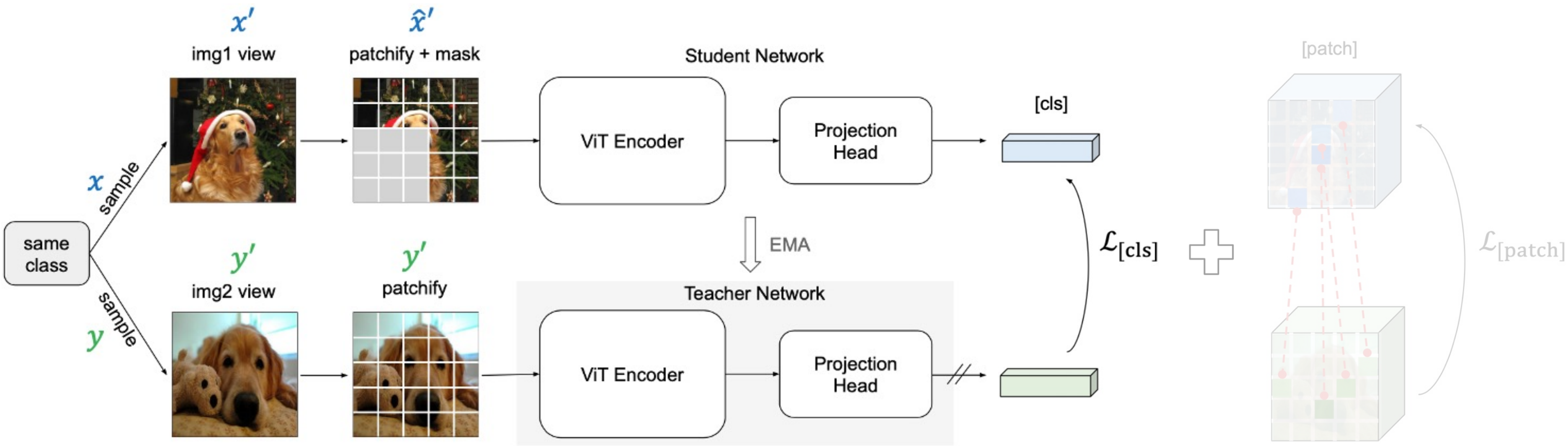


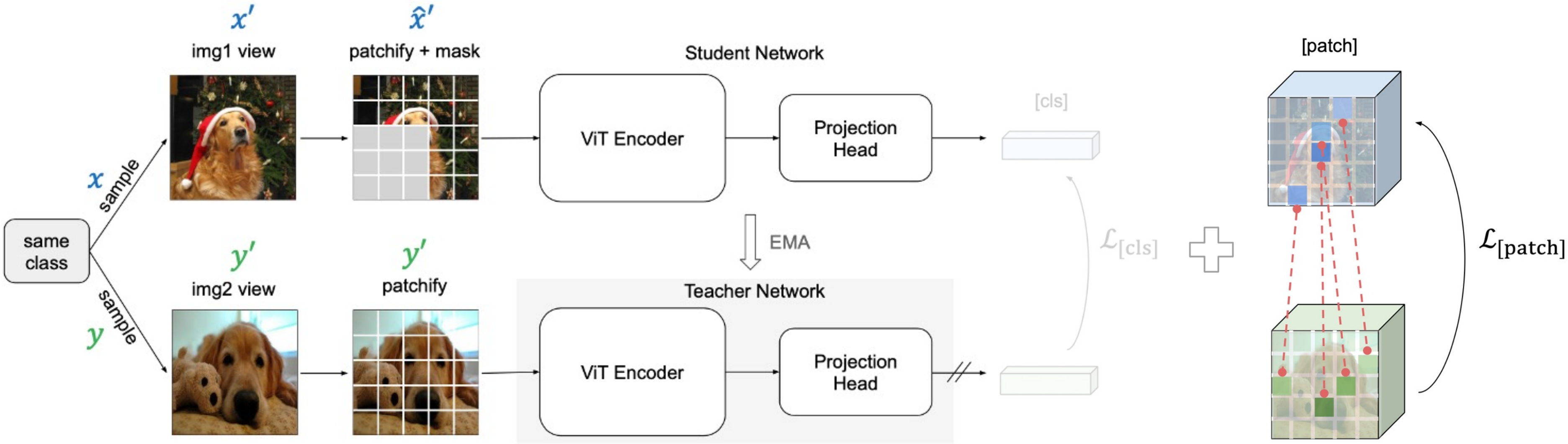
Image source: modified from PAL [4]

Our SMKD Framework (Global Knowledge Distillation)



$$\mathcal{L}_{[\mathrm{cls}]} = -\boldsymbol{P_t}^{[\mathrm{cls}]}(\boldsymbol{y'}) \log \boldsymbol{P_s}^{[\mathrm{cls}]}(\widehat{\boldsymbol{x}}')$$

Our SMKD Framework (Local Knowledge Distillation)

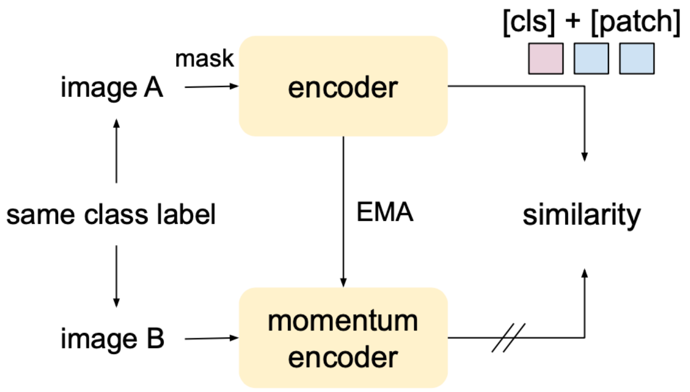


Finding dense correspondence of matched token pairs with highest similarities.

$$\mathcal{L}_{[\mathrm{patch}]} = -\sum_{k=1}^{N} \omega_{k+} \cdot \boldsymbol{P_t}^{[\mathrm{patch}]}(\boldsymbol{y'}_k) \log \boldsymbol{P_s}^{[\mathrm{patch}]}(\widehat{\boldsymbol{x}'}_{k+}) \qquad k^+ = \arg\max_{l \in [N]} \frac{{\boldsymbol{f}_k^t}^\top \boldsymbol{f}_l^s}{\|\boldsymbol{f}_k^t\| \|\boldsymbol{f}_l^s\|}$$

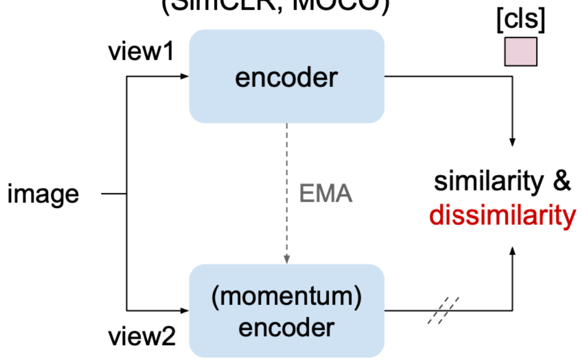
Comparison with Self-Supervised / Supervised Contrastive Frameworks

Our supervised masked knowledge distillation

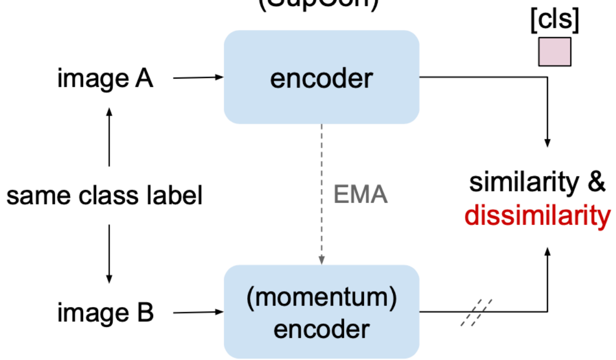


✔ Avoids the need for negative examples ("dissimilarity")

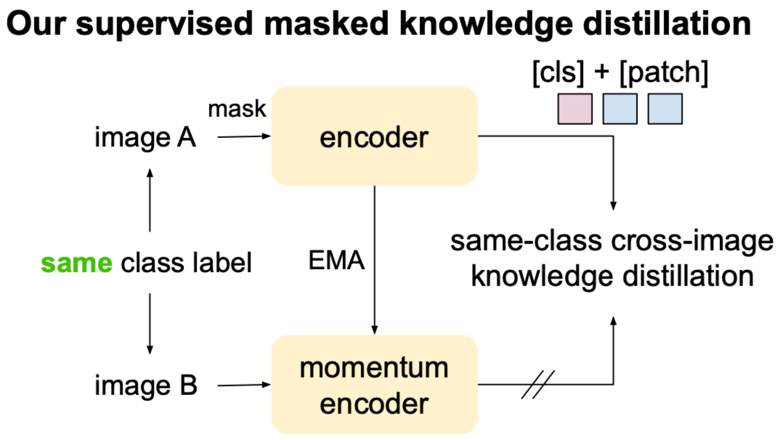
self-supervised contrastive
(SimCLR, MOCO)

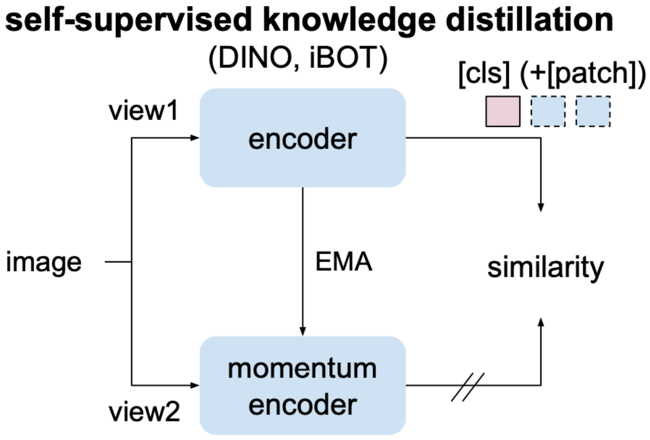


supervised contrastive
(SupCon)



Comparison with Self-Supervised Knowledge Distillation Framework





✔ Introduces **intra-class** knowledge distillation to self-supervised knowledge distillation framework

Dataset Description

- We test on four widely-used few-shot classification benchmark datasets:

	Resolution	#Images	#Classes	#Images per class	(train, val, test) split
CIFAR-FS	32×32	60000	100	600	(64, 16, 20)
FC100	32×32	60000	100	600	(60, 20, 20)
mini-ImageNet	224×224	60000	100	600	(64, 16, 20)
tiered-ImageNet	224×224	779165	608	$\approx$1282	(351, 97, 160)

- Comparison of dataset size:

$$\text{CIFAR-FS} \approx \text{FC100} \leq \textit{mini}\text{-ImageNet} \ll \textit{tiered}\text{-ImageNet}$$

Visualizations

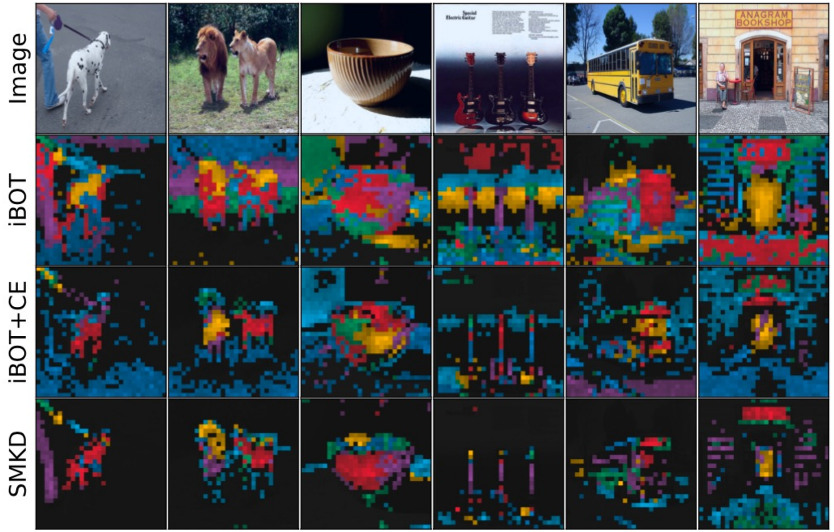


Figure 4. **Visualization of multi-head self-attention maps**. The self-attention of the `[cls]` tokens with different heads in the last attention layer of ViT are visualized in different colors. iBOT+CE represents the model first pre-trained with iBOT, then trained with CE loss. Our SMKD pays more attention to the foreground objects, especially the most discriminative parts.

Figure 5. **Visualization of dense correspondence**. We use the patches with the highest self-attention of the `[cls]` token on each attention head (6 in total) of the last layer of ViT-S as queries. Best-matched patches with the highest similarities are connected with lines.

Comparison with SOTAs

- Our method with simple *Prototype* and *Linear Classifier* evaluation methods could beat all models with CNN backbone.

- Our method grows more effective on datasets with smaller resolutions and fewer training images.

- Our method, combined with tricks from HCT (spectral tokens pooling & small patch size), achieves a new SOTA on *mini*-ImageNet, CIFAR-FS, and FC100, and is comparable with current SOTA (HCTransformers) on *tiered*-ImageNet.

Table 1. **Results on mini-ImageNet and tiered-ImageNet.** Top three methods are colored in red, blue and green respectively.

Method	Backbone	#Params	miniImageNet,5-way		tieredImageNet,5-way	
			1-shot	5-shot	1-shot	5-shot
DeepEMD [73]	*ResNet-12*	12.4M	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
IE [50]	*ResNet-12*	12.4M	67.28 ± 0.80	84.78 ± 0.52	72.21 ± 0.90	87.08 ± 0.58
PAL [41]	*ResNet-12*	12.4M	69.37 ± 0.64	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
DC [71]	*ResNet-12*	12.4M	68.57 ± 0.55	82.88 ± 0.42	78.19 ± 0.25	89.90 ± 0.41
COSOC [40]	*ResNet-12*	12.4M	69.28 ± 0.49	85.16 ± 0.42	73.57 ± 0.43	87.57 ± 0.10
FEAT [72]	*WRN-28-10*	36.5M	65.10 ± 0.20	81.11 ± 0.14	70.41 ± 0.23	84.38 ± 0.16
Meta-QDA [75]	*WRN-28-10*	36.5M	67.38 ± 0.55	84.27 ± 0.75	74.29 ± 0.66	89.41 ± 0.77
OM [46]	*WRN-28-10*	36.5M	66.78 ± 0.30	85.29 ± 0.41	71.54 ± 0.29	87.79 ± 0.46
SUN [16]	*ViT*	12.5M	67.80 ± 0.45	83.25 ± 0.30	72.99 ± 0.50	86.74 ± 0.33
FewTURE [29]	*Swin-Tiny*	29.0M	72.40 ± 0.78	86.38 ± 0.49	76.32 ± 0.87	89.96 ± 0.55
HCTransformers [28]	3×*ViT-S*	63.0M	74.74 ± 0.17	89.19 ± 0.13	79.67 ± 0.20	91.72 ± 0.11
Ours (Prototype)	*ViT-S*	21M	74.28 ± 0.18	88.82 ± 0.09	78.83 ± 0.20	91.02 ± 0.12
Ours (Classifier)	*ViT-S*	21M	74.10 ± 0.17	88.89 ± 0.09	78.81 ± 0.21	91.21 ± 0.11
Ours + HCT [28]	3×*ViT-S*	63M	75.32 ± 0.18	89.57 ± 0.09	79.74 ± 0.20	91.68 ± 0.11

Table 2. **Results on CIFAR-FS and FC100**. Top three methods are colored in red, blue and green respectively.

Method	Backbone	#Params	CIFAR-FS,5-way		FC100,5-way	
			1-shot	5-shot	1-shot	5-shot
BML [79]	*ResNet-12*	12.4M	73.45 ± 0.47	88.04 ± 0.33	45.00 ± 0.41	63.03 ± 0.41
IE [50]	*ResNet-12*	12.4M	77.87 ± 0.85	89.74 ± 0.57	47.76 ± 0.77	65.30 ± 0.76
PAL [41]	*ResNet-12*	12.4M	77.10 ± 0.70	88.00 ± 0.50	47.20 ± 0.60	64.00 ± 0.60
TPMN [66]	*ResNet-12*	12.4M	75.50 ± 0.90	87.20 ± 0.60	46.93 ± 0.71	63.26 ± 0.74
MN+MC [74]	*ResNet-12*	12.4M	74.63 ± 0.91	86.45 ± 0.59	46.40 ± 0.81	61.33 ± 0.71
ConstellationNet [70]	*ResNet-12*	12.4M	75.40 ± 0.20	86.80 ± 0.20	43.80 ± 0.20	59.70 ± 0.20
PSST [11]	*WRN-28-10*	36.5M	77.02 ± 0.38	88.45 ± 0.35	-	-
Meta-QDA [75]	*WRN-28-10*	36.5M	75.95 ± 0.59	88.72 ± 0.79	-	-
SUN [16]	*ViT*	12.5M	78.37 ± 0.46	88.84 ± 0.32	-	-
FewTURE [29]	*Swin-Tiny*	29.0M	77.76 ± 0.81	88.90 ± 0.59	47.68 ± 0.78	63.81 ± 0.75
HCTransformers [28]	3×*ViT-S*	63.0M	78.89 ± 0.18	90.50 ± 0.09	48.27 ± 0.15	66.42 ± 0.16
Ours (Prototype)	*ViT-S*	21M	80.08 ± 0.18	90.63 ± 0.13	50.38 ± 0.16	68.37 ± 0.16
Ours (Classifier)	*ViT-S*	21M	79.82 ± 0.18	90.91 ± 0.13	50.28 ± 0.16	68.50 ± 0.16

Thank you!

Any suggestions and comments are welcome!

Reference

[1] Khosla, Prannay, et al. "Supervised contrastive learning." Advances in Neural Information Processing Systems 33 (2020): 18661-18673.

[2] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[3] Zhou, Jinghao, et al. "Image BERT Pre-training with Online Tokenizer." International Conference on Learning Representations. 2021.

[4] Ma, Jiawei, et al. "Partner-assisted learning for few-shot image classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[5] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[6] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[7] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[8] Chen, Xinlei, Saining Xie, and Kaiming He. "An empirical study of training self-supervised vision transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.