7. Appendix

7.1. Implementation Details for Model Training

Self-supervised pretraining. We pre-train our Vision Transformer backbone and projection head following the same pipeline in iBOT [88]. Most of the hyper-parameter settings are kept unchanged without tuning. Vit-Small, which has ∼21M parameters is used as our default architecture. Our default patch size is set as 16. For the student network, the `[cls]` token output and `[patch]` tokens output share the same projection head. This head-sharing strategy is also applied to the teacher network. For both networks, we set the output dimension of projection heads as 8192. We linearly warm up the learning rate for 10 epochs to its base value of 5e-4, then use cosine schedule to decay it to 1e-5. Cosine schedule is also used for weight decay from 0.04 to 0.4. Besides, we use the multi-crop strategy [6] with 2 global crops (224×224) and 10 local crops (96×96), with scale range $(0.4, 1.0)$ and $(0.05, 0.4)$ respectively. We found that allowing knowledge distillation between global and local crops from intra-class images harms the performance, which is consistent with [88]. Therefore, local crops here are only used for self-distillation with global crops from the same image. Furthermore, we apply blockwise masking on global crops sent into the student network, with a masking ratio uniformly sampled from $[0, 1, 0.5]$ with probability 0.5, and 0 with probability 0.5. Ablation of different masking strategies is given in Sec. 7.3. Our batch size is set as 640 (batch size per GPU equal to 80). *mini*-ImageNet and *tiered*-ImageNet are pre-trained for 1200 epochs, and CIFAR-FS and FC100 are pre-trained for 900 epochs. All models are trained on 8 Nvidia RTX 3090 GPUs.

Supervised knowledge distillation. After finishing the pretraining stage, we train the model with our supervised-contrastive loss. The best evaluation accuracy on the validation set can usually be achieved within 60 epochs of training. We use the same set of hyper-parameters as the first pretraining stage without further tuning. Ablation of the scaling parameter λ, which controls the relative size of $\mathcal{L}_{\texttt{[patch]}}$ and $\mathcal{L}_{\texttt{[cls]}}$ is given in Sec. 7.3.

7.2. Few-shot Evaluation Results

We present few-shot evaluation results with more methods on the four benchmark datasets here in Table 9 and 10. ViT-based methods are better than the traditional CNN-based methods in general. The ranking of our method remains unchanged.

Table 9. **More comprehensive few-shot evaluation 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 [83]	*ResNet-12*	12.4M	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
IE [58]	*ResNet-12*	12.4M	67.28 ± 0.80	84.78 ± 0.52	72.21 ± 0.90	87.08 ± 0.58
BML [89]	*ResNet-12*	12.4M	67.04 ± 0.63	83.63 ± 0.29	68.99 ± 0.50	85.49 ± 0.34
PAL [48]	*ResNet-12*	12.4M	69.37 ± 0.64	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
TPMN [74]	*ResNet-12*	12.4M	67.64 ± 0.63	83.44 ± 0.43	72.24 ± 0.70	86.55 ± 0.63
MN+MC [84]	*ResNet-12*	12.4M	67.14 ± 0.80	83.82 ± 0.51	74.58 ± 0.88	86.73 ± 0.61
DC [78]	*ResNet-12*	12.4M	68.57 ± 0.55	82.88 ± 0.42	78.19 ± 0.25	89.90 ± 0.41
MELR [20]	*ResNet-12*	12.4M	67.40 ± 0.43	83.40 ± 0.28	72.14 ± 0.51	87.01 ± 0.35
COSOC [47]	*ResNet-12*	12.4M	69.28 ± 0.49	85.16 ± 0.42	73.57 ± 0.43	87.57 ± 0.10
CSEI [42]	*ResNet-12*	12.4M	68.94 ± 0.28	85.07 ± 0.50	73.76 ± 0.32	87.83 ± 0.59
CNL [86]	*ResNet-12*	12.4M	67.96 ± 0.98	83.36 ± 0.51	73.42 ± 0.95	87.72 ± 0.75
FEAT [80]	*WRN-28-10*	36.5M	65.10 ± 0.20	81.11 ± 0.14	70.41 ± 0.23	84.38 ± 0.16
Meta-QDA [85]	*WRN-28-10*	36.5M	67.38 ± 0.55	84.27 ± 0.75	74.29 ± 0.66	89.41 ± 0.77
OM [53]	*WRN-28-10*	36.5M	66.78 ± 0.30	85.29 ± 0.41	71.54 ± 0.29	87.79 ± 0.46
SUN [17]	*ViT*	12.5M	67.80 ± 0.45	83.25 ± 0.30	72.99 ± 0.50	86.74 ± 0.33
FewTURE [36]	*ViT-S*	21.0M	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
FewTURE [36]	*Swin-Tiny*	29.0M	72.40 ± 0.78	86.38 ± 0.49	76.32 ± 0.87	89.96 ± 0.55
HCT (Prototype) [79]	3×*ViT-S*	63.0M	74.74 ± 0.17	85.66 ± 0.10	79.67 ± 0.20	89.27 ± 0.13
HCT (Classifier) [79]	3×*ViT-S*	63.0M	74.62 ± 0.20	89.19 ± 0.13	79.57 ± 0.20	91.72 ± 0.11
Ours (Prototype)	*ViT-S*	21.0M	74.28 ± 0.18	88.82 ± 0.09	78.83 ± 0.20	91.02 ± 0.12
Ours (Classifier)	*ViT-S*	21.0M	74.10 ± 0.17	88.89 ± 0.09	78.81 ± 0.21	91.21 ± 0.11
Ours + HCT [79]	3×*ViT-S*	63.0M	75.32 ± 0.18	89.57 ± 0.09	79.74 ± 0.20	91.68 ± 0.11

Table 10. **More comprehensive few-shot evaluation 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
DSN-MR [60]	*ResNet-12*	12.4M	75.60 ± 0.90	86.20 ± 0.60	-	-
BML [89]	*ResNet-12*	12.4M	73.45 ± 0.47	88.04 ± 0.33	45.00 ± 0.41	63.03 ± 0.41
IE [58]	*ResNet-12*	12.4M	77.87 ± 0.85	89.74 ± 0.57	47.76 ± 0.77	65.30 ± 0.76
PAL [48]	*ResNet-12*	12.4M	77.10 ± 0.70	88.00 ± 0.50	47.20 ± 0.60	64.00 ± 0.60
TPMN [74]	*ResNet-12*	12.4M	75.50 ± 0.90	87.20 ± 0.60	46.93 ± 0.71	63.26 ± 0.74
MN+MC [84]	*ResNet-12*	12.4M	74.63 ± 0.91	86.45 ± 0.59	46.40 ± 0.81	61.33 ± 0.71
RENet [37]	*ResNet-12*	12.4M	74.51 ± 0.46	86.60 ± 0.32	-	-
ConstellationNet [77]	*ResNet-12*	12.4M	75.40 ± 0.20	86.80 ± 0.20	43.80 ± 0.20	59.70 ± 0.20
ALFA+MeTAL [2]	*ResNet-12*	12.4M	-	-	44.54 ± 0.50	58.44 ± 0.42
MixtFSL [1]	*ResNet-12*	12.4M	-	-	41.50 ± 0.67	58.39 ± 0.62
CC+rot [23]	*WRN-28-10*	36.5M	73.62 ± 0.31	86.05 ± 0.22	-	-
PSST [12]	*WRN-28-10*	36.5M	77.02 ± 0.38	88.45 ± 0.35	-	-
Meta-QDA [85]	*WRN-28-10*	36.5M	75.95 ± 0.59	88.72 ± 0.79	-	-
SUN [17]	*ViT*	12.5M	78.37 ± 0.46	88.84 ± 0.32	-	-
FewTURE [36]	*ViT-S*	21.0M	76.10 ± 0.88	86.14 ± 0.64	46.20 ± 0.79	63.14 ± 0.73
FewTURE [36]	*Swin-Tiny*	29.0M	77.76 ± 0.81	88.90 ± 0.59	47.68 ± 0.78	63.81 ± 0.75
HCT (Prototype) [79]	3×*ViT-S*	63.0M	78.89 ± 0.18	87.73 ± 0.11	48.27 ± 0.15	61.49 ± 0.15
HCT (Classifier) [79]	3×*ViT-S*	63.0M	78.88 ± 0.18	90.50 ± 0.09	48.15 ± 0.16	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

7.3. Additional Ablation Studies

Why $\mathcal{L}_{[cls]} + \mathcal{L}_{\text{MIM}}$ in stage 1? Our insight is that the `[cls]` tokens in global loss have better high-level semantics, but often disregard the rich local structures. While the MIM loss $\mathcal{L}_{\text{MIM}}$ constructed from `[patch]` tokens can remedy this problem, increase task difficulty, and work as strong data augmentations. In Table 11, we can find that using both losses in stage 1 gives the best results.

Table 11. Ablation of SSL tasks in stage 1 on *mini*-ImageNet.

Stage1				Stage2: $\mathcal{L}_{\texttt{[cls]}} + \mathcal{L}_{\texttt{[patch]}}$	
$\mathcal{L}_{[cls]}$	$\mathcal{L}_{\text{MIM}}$	1-shot	5-shot	1-shot	5-shot
✓		58.55	78.90	72.93	88.07
	✓	27.66	33.82	37.03	50.95
✓	✓	60.93	80.38	74.28	88.82

Masking Strategies. We use blockwise masking as our default in the main text. In Table 12, we test random mask and no mask while keeping all other hyper-parameters unchanged. "Block Mask $\rightarrow$ No Mask" represents self-supervised pretraining with blockwise masking, and supervised training with no mask. Using either a random mask or block mask can boost the classification accuracy in the first self-supervised pretraining stage, but their advantage over no mask decreases in the second supervised training stage. We choose blockwise masking as our default strategy since it balances 1 and 5-shot classification accuracy the best.

Scaling Parameter λ. This parameter controls the relative importance of class-level and patch-level losses in our final loss: $\mathcal{L} = \mathcal{L}_{[\mathbf{cls}]} + \lambda\mathcal{L}_{[\mathbf{patch}]}$. A relatively large value of λ will put more focus on localization and less on high-level semantics. Here in Table 13, we test different λ values by keeping the base of $\mathcal{L}_{[\mathbf{cls}]}$ to 1 and scale $\mathcal{L}_{[\mathbf{patch}]}$. As we can see, the λ parameter influences 1-shot classification accuracy more than 5-shot. We choose $\lambda = 0.25$ as our default (which makes the ratio of $\mathcal{L}_{[\mathbf{patch}]}/\mathcal{L}_{[\mathbf{cls}]}$ roughly around 2) since it has best 1-shot performance and competitive 5-shot accuracy.

Table 12. Ablation over different masking strategy in self-supervised pretraning stage.

Masking Strategy	Self-supervised Pre-train		Supervised Training	
	1-shot	5-shot	1-shot	5-shot
No Mask	59.15 ± 0.17	79.23 ± 0.12	73.94 ± 0.17	88.93 ± 0.09
Block Mask	60.93 ± 0.17	80.38 ± 0.12	74.01 ± 0.17	88.89 ± 0.09
Random Mask	60.94 ± 0.18	79.62 ± 0.13	74.07 ± 0.18	88.66 ± 0.09
Block Mask $\rightarrow$ No Mask	60.93 ± 0.17	80.38 ± 0.12	73.44 ± 0.17	88.87 ± 0.09

Table 13. The influence of different ratio between $\mathcal{L}_{[\mathbf{cls}]}$ and $\mathcal{L}_{[\mathbf{patch}]}$.

$\mathcal{L}_{[\mathbf{patch}]}/\mathcal{L}_{[\mathbf{cls}]}$	1-shot	5-shot
$\approx$4 ($\lambda = 0.9$)	73.45 ± 0.17	88.89 ± 0.09
$\approx$2 ($\lambda = 0.45$)	74.28 ± 0.18	88.82 ± 0.09
$\approx$1 ($\lambda = 0.2$)	74.01 ± 0.17	88.89 ± 0.09
$\approx$0.5 ($\lambda = 0.1$)	73.11 ± 0.17	88.44 ± 0.09

Weighting Parameter ω_{k^+} in $\mathcal{L}_{[\mathbf{patch}]}$. This parameter in Eq.(5) gives weights to each component of our patch-level contrastive loss $\mathcal{L}_{[\mathbf{patch}]}$. We set $\omega_{k^+} = 1/N$ in our main text due to its simplicity. In Table 14, we compare it *(Simple Avg)* with another variant *(Self-Attention Weighted Avg)*, which uses the averaged self-attention weights over attention heads of the `[cls]` token with all `[patch]` tokens in the last attention layer of teacher network to aggregate pairwise patch matching losses. As found in [7], the self-attention of ViTs is good at capturing foreground regions. So we use it here as a way to highlight foreground objects and to attenuate irrelevant background information. Our default simple average outperforms this variant on both 1 and 5-shot classification accuracies. One explanation is as follows. If the foreground objects of two intra-class images differ a lot, then *Self-Attention Weighted Avg* tends to minimize $\mathcal{L}_{[\mathbf{patch}]}$ by decreasing the weights of the losses associated with the patches covering these foreground objects, which makes our model deviate from optimum.

Table 14. Different weighting schemes of patch-level supervised-contrastive loss.

Weighting Scheme	1-shot	5-shot
Simple Avg	74.28 ± 0.18	88.82 ± 0.09
Self-Attention Weighted Avg	74.11 ± 0.18	88.52 ± 0.10

Comparison with smaller backbones: To make ViT-S ($\sim$ 21M parameters) comparable with ResNet-12 ($\sim$ 12M parameters), we trim by half either its embedding dimension (d_{embed}) or the number of attention heads (#heads). From Table 15, trimming #heads by half only results in little drop in accuracy, which still outperforms the best method with ResNet-12 backbones. The training speed also increases by 10% with fewer #heads. Given this result, our comparison now becomes complete: our method outperforms both shallow (ResNet-12) and deep (WRN-28-10) CNN-based backbones, as well as ViTs with the same (see Table 5) or more (see Table 3 & Table 4) parameters.

Table 15. ViT-S with similar size as ResNet-12 on *mini*-ImageNet

Backbone	d_{embed}	#heads	#param	Stage1		Stage2	
				1-shot	5-shot	1-shot	5-shot
ViT-S	**192**	6	11M	60.70	79.56	71.14	87.12
ViT-S	384	**3**	11M	62.12	81.27	**72.70**	**87.90**
ViT-S	384	6	21M	60.93	80.38	74.28	88.82
ResNet12	-	-	12M	-	-	69.37	85.16
WRN-28-10	-	-	36M	-	-	67.38	85.29

7.4. Visualizations

We visualize more self-attention maps and dense correspondence in Fig. 6 and Fig. 7.

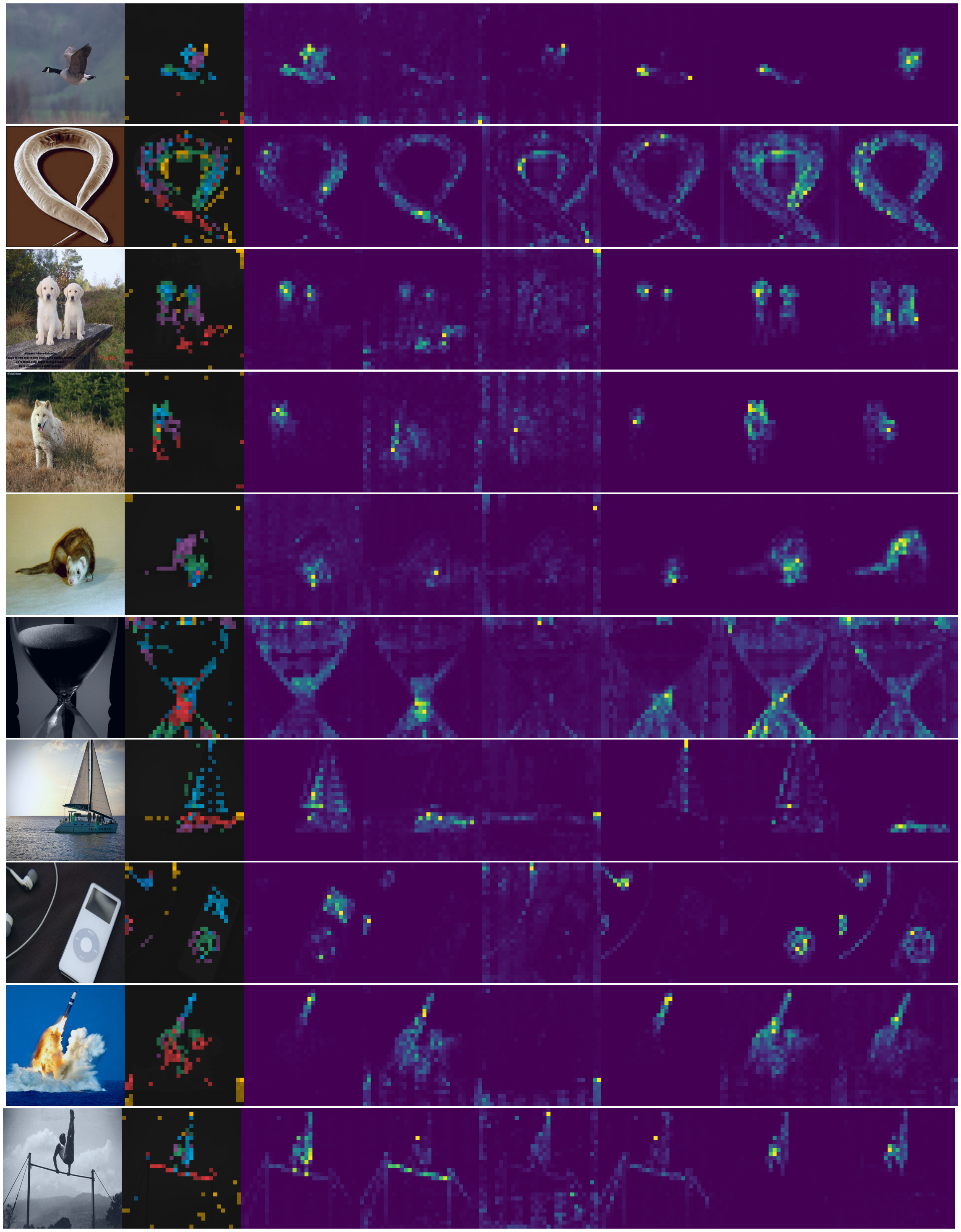

Figure 6. **Visualization of multi-head self-attention maps**. The self-attention of the `[cls]` tokens with different heads in the last attention layer of ViT are visualized in different colors in the second column. The last six columns visualize each attention head. Images are from the test set of *mini*-ImageNet.

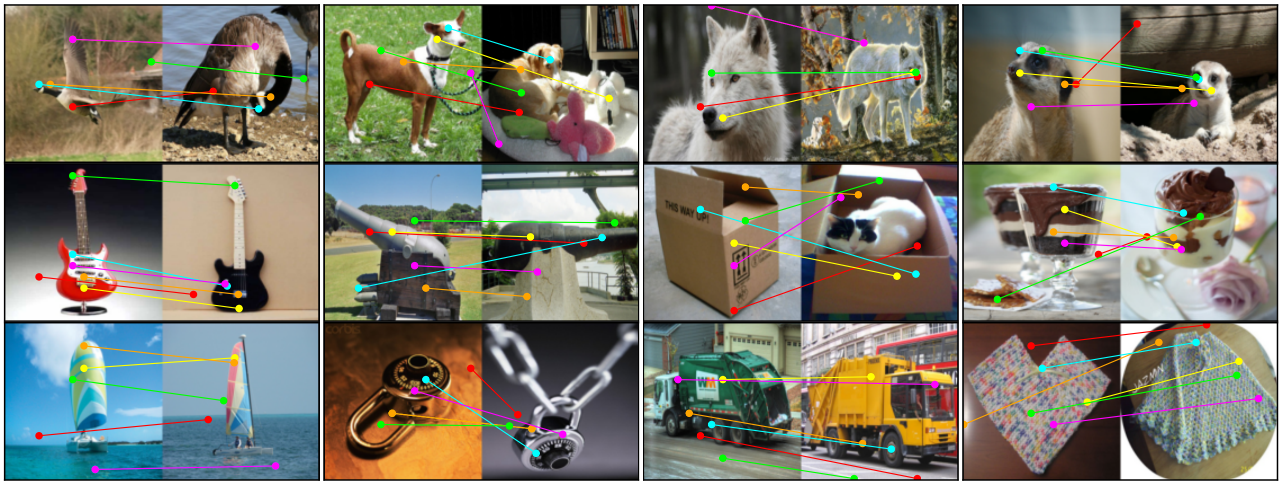

Figure 7. **Visualization of dense correspondence**. We use the patches with the highest self-attention of the `[cls]` token on each attention head (6 in total) of the last layer of ViT-S as queries. Best-matched patches with the highest similarities are connected with lines. Images are from the validation and testing set of *mini*-ImageNet.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Mixture-based feature space learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9041–9051, 2021. 14

[2] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9465–9474, 2021. 14

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3, 4

[4] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018. 5

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 13

[7] 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, 3, 4, 15

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[9] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 3

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 4

[11] 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. 2

[12] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. Pareto self-supervised training for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13663–13672, 2021. 6, 14

[13] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 1

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993, 2020. 3

[17] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Self-promoted supervision for few-shot transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 6, 13, 14

[18] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 2

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[20] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. Melr: Meta-learning via modeling episode-level relationships for few-shot learning. In *International Conference on Learning Representations*, 2020. 13

[21] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. *arXiv preprint arXiv:2110.06014*, 2021. 3

[22] Hanan Gani, Muzammal Naseer, and Mohammad Yaqub. How to train vision transformer on small-scale datasets? *arXiv preprint arXiv:2210.07240*, 2022. 2, 3

[23] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8059–8068, 2019. 1, 14

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4

[25] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pages 124–141. Springer, 2020. 3

[26] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3263–3272, October 2021. 3

[27] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789, 2022. 3

[28] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5321–5330, June 2022. 3

[29] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, Rama Chellappa, and Shih-Fu Chang. Multimodal few-shot object detection with meta-learning based cross-modal prompting. *arXiv preprint arXiv:2204.07841*, 2022. 3

[30] Guangxing Han, Xuan Zhang, and Chongrong Li. Revisiting faster r-cnn: A deeper look at region proposal network. In *International Conference on Neural Information Processing*, pages 14–24, 2017. 1

[31] Guangxing Han, Xuan Zhang, and Chongrong Li. Single shot object detection with top-down refinement. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3360–3364. IEEE, 2017. 1

[32] Guangxing Han, Xuan Zhang, and Chongrong Li. Semi-supervised dff: Decoupling detection and feature flow for video object detectors. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1811–1819, 2018. 1

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[36] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. *arXiv preprint arXiv:2206.07267*, 2022. 2, 3, 6, 7, 13, 14

[37] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8822–8833, 2021. 14

[38] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[41] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021. 1

[42] Junjie Li, Zilei Wang, and Xiaoming Hu. Learning intact features by erasing-inpainting for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8401–8409, 2021. 13

[43] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021. 1, 2, 3

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 3

[45] Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang. Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653*, 2020. 1

[46] Yuning Lu, Liangjian Wen, Jianzhuang Liu, Yajing Liu, and Xinmei Tian. Self-supervision can be a good few-shot learner. *arXiv preprint arXiv:2207.09176*, 2022. 2

[47] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background: Shared object concentration for few-shot image recognition. *arXiv preprint arXiv:2107.07746*, 2021. 6, 13

[48] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10573–10582, 2021. 1, 3, 4, 13, 14

[49] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2218–2227, 2020. 2, 3, 5

[50] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018. 3, 5

[51] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. A unified view of masked image modeling. *arXiv preprint arXiv:2210.10615*, 2022. 2, 3

[52] Elia Peruzzo, Enver Sangineto, Yahui Liu, Marco De Nadai, Wei Bi, Bruno Lepri, and Nicu Sebe. Spatial entropy regularization for vision transformers. *arXiv preprint arXiv:2206.04636*, 2022. 2, 3

[53] Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8412–8422, 2021. 6, 13

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[55] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[56] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1

[57] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 5

[58] Mamshad Nayeem Rizve, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10836–10846, June 2021. 6, 13, 14

[59] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. 3

[60] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020. 14

[61] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 3, 5

[62] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 1

[63] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 3

[64] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021. 4

[65] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020. 3

[66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 3

[67] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 8

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[69] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 1, 3, 5

[70] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 3

[71] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 3

[72] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 5

[73] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1, 3

[74] Jiamin Wu, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8433–8442, 2021. 6, 13, 14

[75] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 2

[76] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 8

[77] Weijian Xu, Yifan Xu, Huaijin Wang, and Zhuowen Tu. Attentional constellation nets for few-shot learning. In *International Conference on Learning Representations*, 2021. 14

[78] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021. 13

[79] Dongyang Zhao Hong-Yu Zhou Weifeng Ge Yizhou Yu Wenqiang Zhang Yangji He, Weihan Liang. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning, 2022. 1, 2, 3, 6, 7, 13, 14

[80] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020. 3, 13

[81] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[82] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 4

[83] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 13

[84] Chi Zhang, Henghui Ding, Guosheng Lin, Ruibo Li, Changhu Wang, and Chunhua Shen. Meta navigator: Search for a good adaptation policy for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9435–9444, 2021. 13, 14

[85] Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. Shallow bayesian meta learning for real-world few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 651–660, 2021. 6, 13, 14

[86] Jiabao Zhao, Yifan Yang, Xin Lin, Jing Yang, and Liang He. Looking wider for better adaptive representation in few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10981–10989, 2021. 13

[87] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020. 4

[88] 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, 3, 4, 5, 13

[89] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8402–8411, 2021. 6, 13, 14

[90] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1