A Appendix / supplemental material

A.1 Dataset Description
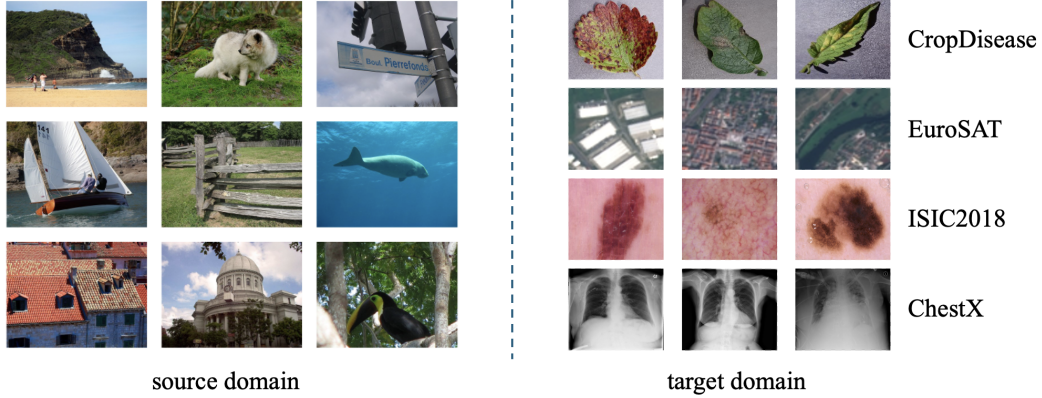
Figure 7: Samples of source domain *mini*Imagenet dataset (left) and target domain datasets (right), from top to bottom correspond to CropDiseases, EuroSAT, ISIC2018, and ChestX. We can see large domain gaps between source and target domains.

***mini*ImageNet** [40] is a subset derived from the larger ImageNet [6] dataset. It consists of 100 categories, each containing 600 natural images. Following the current works [2, 12], we employ the training set of *mini*ImageNet as the source domain dataset, consisting of 64 classes and 38,400 images. Additionally, as depicted in Fig. 7, we utilize datasets from four distinct domains as target domains following [12], including plant disease images, surface satellite imagery, skin disease images, and chest X-ray images. We will introduce them sequentially below.

CropDiseases [30] is a dataset for agricultural disease recognition, encompassing 38 distinct classes and a total of 43,456 images. The dataset comprises images of various crops, including infected and healthy plants, and corresponding disease category labels.

EuroSAT [13] is a comprehensive dataset comprising satellite imagery of the Earth. It encompasses a total of 27,000 images distributed across 10 distinct classes. The dataset offers a diverse range of geographical and topographical features.

ISIC2018 [5] is a medical imaging dataset for skin lesion classification. The dataset consists of 10,015 images categorized into 7 distinct classes.

ChestX [44] is an X-ray medical imaging dataset for chest classification. The dataset consists of 25,847 images across 7 different classes.

A.2 More experiments

A.2.1 Applying Our Method to ViT Variants

Table 7: Our method with iBOT-pretrained ViT-S.

Method	Shot	ChestX	ISIC2018	EuroSAT	CropDiseases	Average
iBOT	1	22.67	31.61	72.85	81.35	52.12
iBOT+Ours	1	**23.01**	**34.69**	**73.04**	**82.39**	**53.28**
iBOT	5	26.31	44.54	89.65	94.79	63.82
iBOT+Ours	5	**27.63**	**51.06**	89.21	**95.20**	**65.78**

We also apply our method to ViT-Small pretrained by iBOT[51] and ViT-Base pretrained by DINO[1]. iBOT is a self-supervised pre-training framework that learns semantic representations of images

15

Table 8: Our method with DINO-pretrained ViT-Base.

Method	Shot	ChestX	ISIC2018	EuroSAT	CropDiseases	Average
DINO-B	1	22.78	34.08	71.89	82.77	52.88
DINO-B+Ours	1	**22.81**	**35.98**	**72.98**	**83.14**	**53.73**
DINO-B	5	26.52	48.77	89.67	95.45	65.10
DINO-B+Ours	5	**27.09**	**52.71**	**90.06**	**95.70**	**66.39**

through Masked Image Modeling (MIM) and an online Tokenizer, enabling effective pre-training of vision Transformers without needing labeled data. The results are shown in Tab. 7. iBOT represents the iBOT-pretrained ViT baseline, while iBOT+Ours denotes our method applied to iBOT. As we can see, our method also shows considerable improvement on the ViT pre-trained with iBOT, with a 1.16 point increase in 1-shot and a 1.96 point increase in 5-shot. The results of our method applied to DINO-pretrained ViT-Base are shown in Tab. 8. DINO-B represents the DINO-pretrained ViT-Base baseline in our CDFSL task, and DINO-B+Ours denotes our method applied to DINO-pretrained ViT-Base. This also shows an improvement in ViT-Base with a 0.85 point increase in 1-shot and a 1.29 point increase in 5-shot.

A.2.2 The Effectiveness of Our Attention
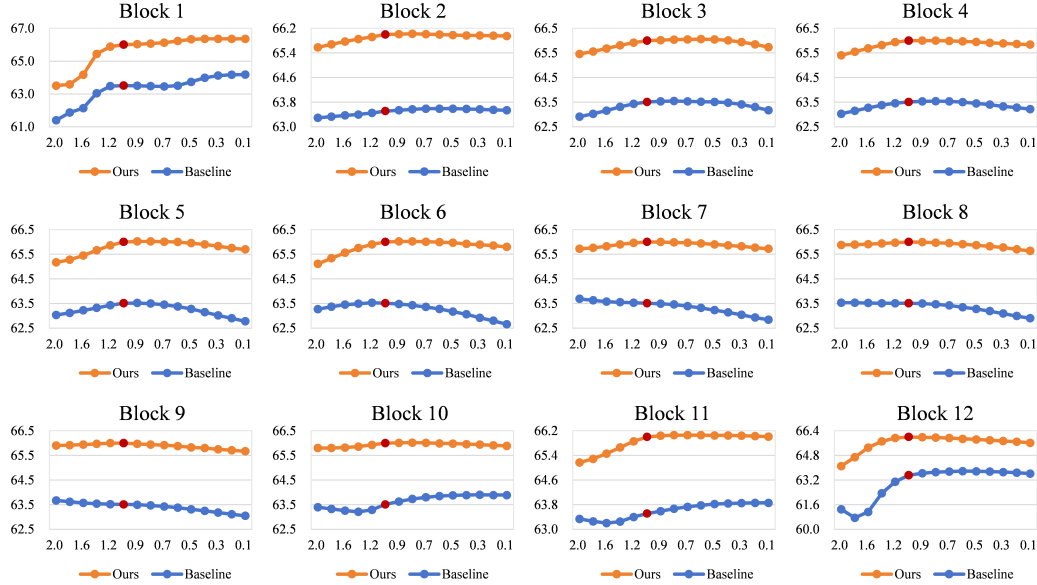
Figure 8: Average target domain accuracy vs. temperature. The red point represents the temperature is 1.0. If the attention is good enough, the attention adjustment will be trivial. Our method shows less reliance on the attention adjustments compared with the baseline method and achieves the best performance when the temperature is 1.0 in most blocks. This indicates the attention produced by our method is improved.

In Fig. 8, we illustrate the average target domain accuracy of the baseline method and our model with attention adjustment. The red point means the temperature is 1.0 (i.e., no temperature is applied). The accuracy of the baseline method increases with the temperature decrease from 2.0 to 0.0 in most blocks, which means temperature adjustment is important in the target domain. If the attention is already good enough, the impact of attention adjustment is trivial. As we can see, our method exhibits little variation in performance with temperature adjustment compared to the baseline method, and the highest accuracy is achieved when the temperature is 1.0 in most blocks. This verifies the effectiveness of our attention.

16

A.3 Broader Impact

Our research introduces an improved ViT model based on attention temperature adjustment, aimed at addressing the ineffective target-domain attention caused by the query-key attention mechanism in CDFSL tasks. By suppressing the learning of query-key parameters and encouraging that of non-query-key parameters, our method significantly enhances the model's cross-domain transferability. This work is not only applicable to CDFSL but can also be extended to other domains, such as domain generalization, domain adaption, and few-shot class-incremental learning, where enhancing the transferability of self-attention is a prevalent challenge. While our method has been evaluated across four distinct target domains, providing a good initial assessment of our method's cross-domain applicability, the diversity of these domains may not encompass all possible real-world scenarios. Future work will aim to extend our evaluations to include a wider range of target domains to understand their performance in diverse real-world settings better.

17