A Appendix / supplemental material

A.1 Detailed Dataset Description
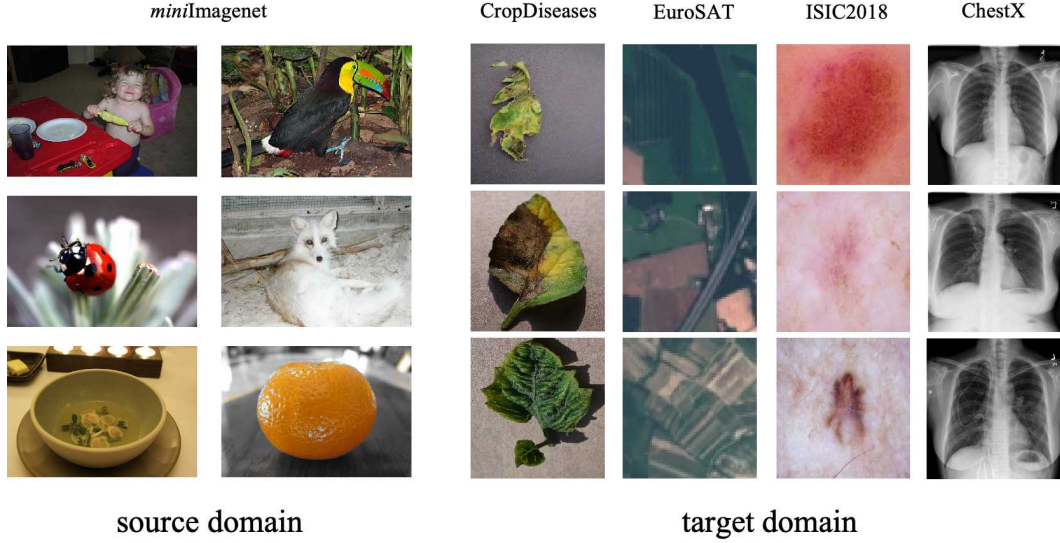
Figure 6: Samples of source domain *mini*ImageNet dataset(left) and target domain datasets (right), from left to right correspond to five distinct datasets: *mini*ImageNet, CropDiseases, EuroSAT, ISIC2018, and ChestX.

***mini*ImageNet**[34] is a subset of the ImageNet[8] dataset that contains 100 categories, each consisting of 600 natural images. Following the current work[28, 33], we split the *mini*ImageNet into 64 classes as the source domain training dataset. In addition, as shown in Figure 6, we utilize the datasets from four other different domains, like agriculture, remote sensing, and medical data, as target domains. We'll sequentially introduce each of them below.

CropDiseases [25] consists of 38 distinct classes and a total of 43,456 images, which are natural images, but are very specialized (specific to the agriculture industry), including various infected crops, healthy plants, and their corresponding disease category labels.

EuroSAT [16] contains a total of 27,000 satellite images of the Earth categorized into 10 distinct classes. The images in the EuroSAT are less similar to images in *mini*ImageNet since they lack perspective distortion, but still color images of natural scenes.

ISIC2018 [5], which is even less similar to the *mini*ImageNet as they could not even represent natural scenes, encompasses 10,015 medical images for skin lesion classification across 7 different classes.

ChestX [38], a medical dataset for chest classification, consists of 25,847 images distributed across 7 distinct classes. The dataset is the most dissimilar to the *mini*ImageNet in three criteria. Apart from the two factors mentioned above, it loses 2 color channels that appear in the ChestX.

A.2 More Experiments

A.2.1 More Visualization of the CLS token by retrieval

To delve into the information encoded by the CLS token, we calculate the cosine similarity between the CLS token and the image token in the input layer of a fixed ViT trained on the source dataset and report the similarity map in Figure 7. It is clear that on the source dataset, the background regions have a higher similarity map, roughly representing the contour of the object for the source domain dataset. It means that the CLS token can distinguish the background and the foreground of the image which could indeed facilitate source-domain recognition. However, as shown in Figure 7 from above to bottom, with images that are less similar to the source domain, the ability to recognize the background for the CLS token gradually fades into mediocrity.

14
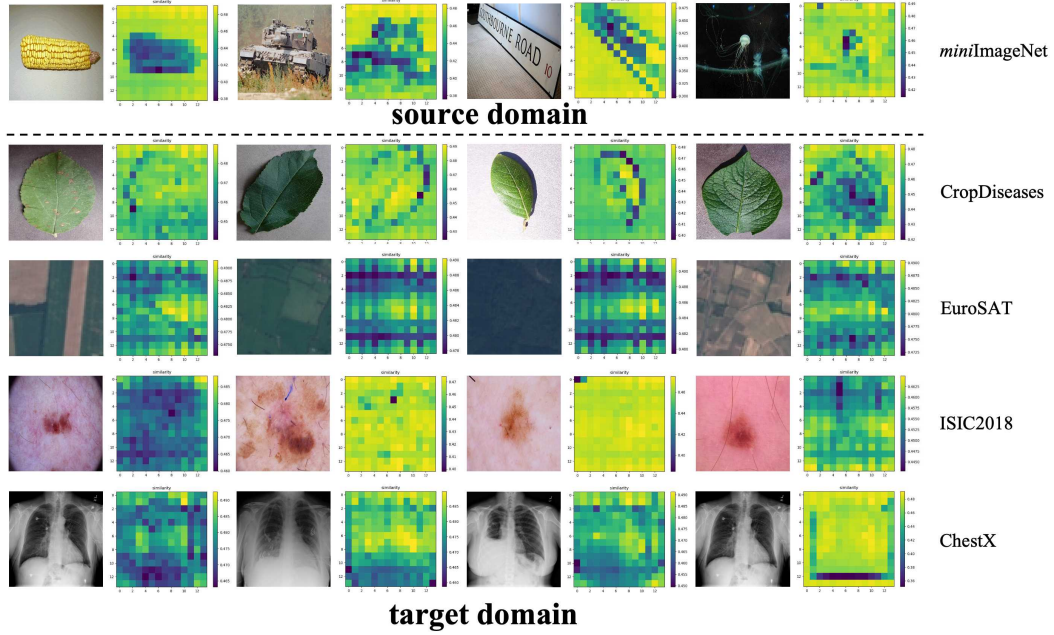
Figure 7: Samples from top to bottom correspond to the similarity map between the CLS token and the image token on the *mini*ImageNet and four target domain datasets, like CropDiseases, EuroSAT, ISIC2018, and ChestX. It can be seen that the CLS token can distinguish the background and the foreground of the image which could indeed facilitate the source-domain recognition on the source dataset while meeting more difficulties with a larger domain gap.

A.2.2 More Visualization of the domain tokens by retrieval
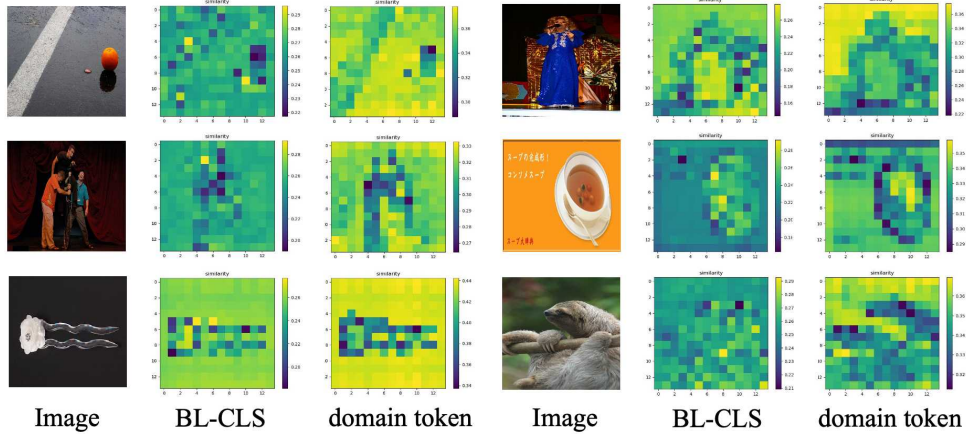
Figure 8: Applying the domain token significantly improves the domain similarity compared to the CLS token of the baseline method (BL-CLS), validating the effectiveness of our approach in absorbing domain information

To delve into what the domain token encodes, we re-utilize the similarity map plotted in Figure 8 as our measurement to compare our approach with the baseline's CLS token (denoted as BL-CLS). It seems that when utilizing the domain token, the domain similarity significantly increases, proving

15

that the domain token can better absorb the domain information and efficiently help to decouple such information from the CLS token.

A.2.3 Applying Our Method to Other Baselines

Table 7: Ablation study of our method with iBOT-pretrained ViT and DINO-Pretrained ViT-Base by the 5-way 1-shot accuracy.

Method	CropDiseases	EuroSAT	ISIC2018	ChestX	Ave.
iBOT	81.17	72.71	31.44	22.56	51.97
iBOT + Ours	**81.31**	**72.80**	**31.87**	**22.57**	**52.14**
DINO-ViT-Base	82.97	72.06	34.19	22.60	52.95
DINO-ViT-Base + Ours	**83.11**	**73.77**	**34.75**	**22.98**	**53.65**

Table 8: Implementing our method with meta-learning baseline.

Method	Crop.	Euro.	ISIC.	Ches.	Ave.
ProtoNet	93.59	86.92	46.15	25.68	63.09
ProtoNet + Ours	**95.03**	**89.42**	**48.67**	**27.15**	**65.07**

We also implement our approach on distinct backbones, like ViT pretrained by iBOT [45], and ViT-Base [42] pretrained by DINO. The results can be seen in Tab 7. Specifically, iBOT represents the iBOT-pretrained ViT baseline, and DINO-ViT-Base corresponds to the ViT-Base pretrained by DINO baseline. It is clarified from the average performance of four target domains that our approach shows improvements among both backbones in the 5-way 1-shot setting.

To verify our model also suits the meta-learning-based baselines, we conduct experiments based on the ProtoNet [30] in Tab 8. We can see that our model also improves this kind of baseline method.

A.2.4 Further Verification of the CLS Token

(a) Visualization of random CLS tokens   (b) Act. Ratio vs. random or CLS token
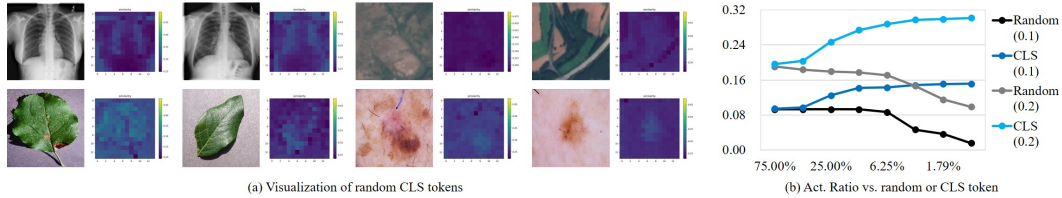
Figure 9: To verify it is the CLS token that tends to capture low-frequency components, we (a) visualize similarity maps of a random token and (b) use random tokens to calculate the similarity map of low-frequency images, and find random tokens do not show the same results as the CLS token.

To verify it is the CLS token that tends to capture low-frequency components, we visualize similarity maps of a random token in Fig. 9a. We use the same color bar as in Fig.2b. We can see the similarity is much lower, but a coarse contour of objects can still be observed in both the source and target domains, indicating a good transferability of detecting object contours, because no domain information is in the random token. However, the contour detected by the random token is much worse than that by the CLS tokens as shown in Fig.2a and Fig.5a, indicating although the random CLS token can initially detect the object contour, the learning of the CLS token strengthens this characteristic to detect low-frequency images.

Then, we use random tokens to calculate the similarity map of low-frequency images, and find random tokens do not show the same results as the CLS token. Specifically, we measure the activation ratio of the CLS token and the random token. We take the top 10% or 20% value as examples in Fig. 9b. We can see the random tokens show a tendency to decrease the activation ratio, while the CLS token shows a tendency to increase the ratio, indicating it is the CLS token that tends to be similar to the feature of low-frequency images.

16

Table 9: Verification of re-initializing the CLS token.

Method	Crop.	Euro.	ISIC.	Ches.	Ave.
Baseline	95.62	88.62	46.08	26.25	63.89
Ours	95.55	90.48	49.58	27.03	65.66
Ours + Re-Init	95.20	89.60	50.23	26.79	65.31

A.2.5 Further Verification of Domain Tokens

To verify that our model does not overfit the fixed random CLS token (so that view the fixed value as a kind of new domain token), we re-initialize the CLS token during the source-domain training. The results are reported in Tab. 9, and we can see the improvements also exist.

However, re-initializing the CLS token can be viewed as adding noise to the domain token, therefore harming the absorbed domain information, which then affects the domain-irrelevant information learned by other structures in ViT. As a result, the Re-Init performance is slightly lower.

Indeed, domain tokens are encouraged to be orthogonal to the fixed CLS token, which would drive the model to view the fixed token as a domain-agnostic token. But note that the random token is already agnostic enough to every domain even without training, therefore our training would not essentially drive the model to be more agnostic to that CLS token, i.e., our model is not bound to the specific value of the CLS token.

Table 10: Training with datasets of 5 constructed domains.

Method	Crop.	Euro.	ISIC.	Ches.	Ave.
Baseline	93.85	89.72	49.74	26.07	64.77
320 classes as domain token	95.03	90.49	49.10	26.81	65.35
64 classes as domain token	95.16	90.61	47.86	26.77	65.10
5 domains as domain token	**95.52**	**90.62**	**50.77**	**27.00**	**65.98**

To further ablate "domain-specific" and "class-specific", we then manually construct some new source domains based on miniImageNet. Specifically, we take the amplitude (by Fourier transformation) from target domains as the style information, and use the phase (by Fourier transformation) from the original source-domain images as the content information, thereby constructing 4 new domains with the original 64 source-domain classes. Then, we train our model on a new dataset containing the 4 constructed datasets and the original source dataset, and ablate different choice of domain tokens in Tab. 10.

As can be seen, by introducing larger domain gaps, viewing each class as a domain is not the best choice. Instead, setting a domain token for each domain could achieve the best performance, which validates the rationale of the domain token in absorbing the domain information.

A.3 Broader Impact

We propose a CD-FSL method based on decoupling the CLS token into domain-specific and domain-agnostic tokens in the ViT and making use of it for efficient downstream few-shot learning to alleviate the domain gap and generalize well to the target domain. This work can also be adopted in other fields, like domain generalization, domain adaption, and few-shot class-incremental learning, where the challenge of enhancing the transferability of model exists universally. The evaluations of our approach are mainly across four different target domains, which may not represent all possible real-world scenarios. The approach can be evaluated on various target domains to be validated in a more realistic setting.

17