Interpretable Cross-Domain Few-Shot Learning with Rectified Target-Domain Local Alignment

Supplementary Material

6. Detailed Dataset Description

Our experimental setup follows the BSCD-FSL [12] benchmark, addressing the challenge of significant distributional shifts across four distinct target domain datasets. Detailed information on these datasets is provided below:

CropDiseases [27] is a dataset including 54,306 images of 14 crop species (Apple, Blueberry, Cherry, Corn, Grape, Orange, Peach, Bell Pepper, Potato, Raspberry, Soybean, Squash, Strawberry, and Tomato) with 26 diseases (or healthy). The samples of this dataset are listed in Fig. 11. CropDiseases images are natural images, but are very specialized (specific to the agriculture industry), so the domain gap here is larger than in the previous cross-domain setting [33].

Figure 11. Samples from CropDiseases.

EuroSAT [14] is a dataset for land use and land cover classification. EuroSAT based on Sentinel-2 satellite imagery, covers 13 spectral bands and consists of 10 categories including Industrial Buildings, Residential Buildings, Annual Crop, Permanent Crop, River, Sea & Lake, Herbaceous Vegetation, Highway, Pasture and Forest, with a total of 27,000 annotated and geographically referenced images. Compared to CropDiseases, EuroSAT images are less similar to *mini*Imagenet (source domain dataset) as they have lost perspective distortion, but are still color images of natural scenes. The samples of this dataset are listed in Fig. 12.

Figure 12. Samples from EuroSAT.

The **ISIC2018** [6] dataset was published by the International Skin Imaging Collaboration (ISIC) as a large-scale dataset of dermoscopy images containing 10,015 images of seven skin injury types (melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, or a vascular lesion). ISIC2018 images are even less similar to *mini*Imagenet as they have lost perspective distortion and no longer represent natural scenes. The samples of this dataset are listed in Fig. 13.

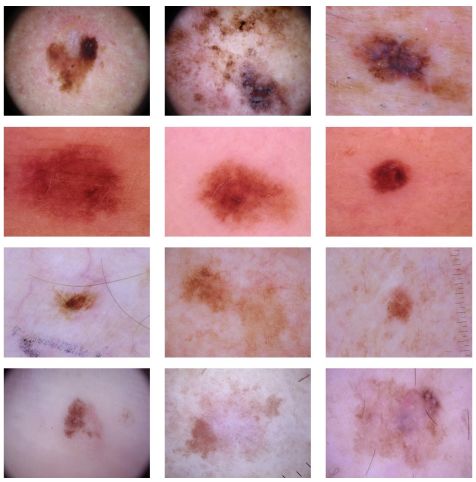

Figure 13. Samples from ISIC2018.

ChestX-ray14 is the largest lung X-ray database to date,

which contains more than 100,000 pre-X-ray views for 14 lung diseases. Categories 1 to 14 correspond to 14 lung diseases, and category 15 indicates no disease. [35] studied the images of eight diseases in this database and constructed the ChestX-ray8 dataset, which comprises 108,948 frontal view X-ray images of 32,717 unique patients with the text-mined eight disease image labels (where each image can have multi-labels) from the associated radiological reports using natural language processing. In this work, we use **ChestX-ray8** for cross-domain testing, consistent with [12]. ChestX is the most dissimilar to *mini*Imagenet across the four target domains as its images have lost perspective distortion, do not represent natural scenes, and have lost 2 color channels. The samples of this dataset are listed in Fig. 14.

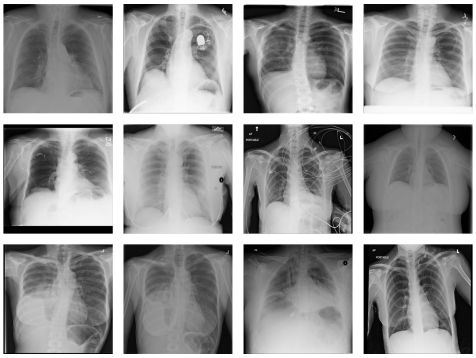

Figure 14. Samples from ChestX.

7. Applying Our Method to other ViT Variants

Table 5 presents a comparative analysis of our method against the CLIP-LoRA baseline across other ViT architectures, specifically ViT-B/32, and ViT-L/14, under the 1-shot setting. As observed, our method consistently demonstrates significant performance improvements over the baseline across all tested ViT variants and target domains.

Table 5. Comparison with ViT-based methods in 1-shot.

Backbone	Method	ChestX	ISIC	EuroSAT	Crop.	Ave.
ViT-B/32	Baseline	21.23	36.32	79.79	81.13	54.62
	+ Ours	21.36	38.26	82.88	83.98	56.62
ViT-L/14	Baseline	22.90	38.45	85.83	89.83	59.25
	+ Ours	23.19	38.72	90.11	90.77	60.70

8. Experimental Data of Figure 2

Figure 2 in the main paper quantifies the hypothesis that the domain gap and scarce training data in CDFSL hurt patch-level local alignment more severely than global alignment. Table 6 reports these alignment scores for both global (CLS-token) and local (patch-level) representations of the baseline, CLIP-LoRA [41]. As observed, all cross-domain datasets exhibit lower local alignment scores compared to miniImageNet, with the exception of EuroSAT, whose images are dominated by foreground regions, confirming the baseline's inability to capture fine-grained cues under data scarcity. To further validate the efficacy of the proposed CC-CDFSL framework, we compare the local alignment scores before and after applying our method. Table 7 summarizes the results. Consistently, our method significantly improves these scores, demonstrating that the cycle-consistency and semantic anchor mechanisms successfully enhance patch-level semantics alignment under domain shift without additional annotations.

Table 6. Alignment Scores of Global and Local Features.

Dataset	Global Feature	Local Features	Far Domain
miniImageNet	0.7056	0.3821	✗
ChestX	0.6768	0.2391	✓
ISIC2018	0.7261	0.2360	✓
EuroSAT	0.6768	0.3845	✓
CropDiseases	0.7012	0.3811	✓

Table 7. Comparison of Local Feature Alignment Scores.

Method	ChestX	ISIC	EuroSAT	Crop.	Ave.
Baseline	0.2391	0.2360	0.3845	0.3811	0.3102
+ Ours	0.3293	0.2661	0.3796	0.3877	0.3407

9. Generalization to Base-to-New Setting

The base-to-new generalization task refers to a scenario where a dataset is evenly divided by class into non-overlapping base classes and new classes. After a model is trained on a few-shot dataset from the base classes, its generalization ability is tested on the new classes. The harmonic mean (HM) of the classification accuracies on both is used to evaluate the overall performance:

$$\mathrm{HM} = \frac{2}{\frac{1}{\mathrm{Base}} + \frac{1}{\mathrm{New}}} = \frac{2 \cdot \mathrm{Base} \cdot \mathrm{New}}{\mathrm{Base} + \mathrm{New}} \tag{17}$$

As shown in Table 8, CC-CDFSL consistently outperforms MaPLe [20] across all datasets in terms of new-class accuracy and harmonic mean, with marginal improvements on base classes.

9.1. Better Class Separation

As illustrated in the Figure 15, features extracted by CLIP-LoRA tend to form less compact and more overlapping clusters, with different classes not well separated in the embedding space. In contrast, our method produces more distinct and compact clusters for each class, resulting in clearer class boundaries and reduced intra-class variance. This indicates that our approach achieves better class separation

Table 8. Base-to-new generalization on 11 datasets. Ours (CC-CDFSL) consistently improves new-class accuracy and harmonic mean (HM) over MaPLe, validating the cross-task generality of cycle-consistent patch-level alignment.

(a) Average over 11 datasets.

	Base	New	HM
MaPLe	82.19	74.55	78.18
+ Ours	82.26	75.75	78.87

(b) ImageNet.

	Base	New	HM
MaPLe	75.57	70.88	73.15
+ Ours	75.59	71.10	73.28

(c) Caltech101.

	Base	New	HM
MaPLe	98.02	94.43	96.19
+ Ours	97.91	95.49	96.68

(d) OxfordPets.

	Base	New	HM
MaPLe	95.64	97.80	96.71
+ Ours	95.82	97.99	96.89

(e) StanfordCars.

	Base	New	HM
MaPLe	72.56	73.80	73.17
+ Ours	72.66	74.84	73.73

(f) Flowers102.

	Base	New	HM
MaPLe	96.11	72.44	82.61
+ Ours	96.07	73.90	83.19

(g) Food101.

	Base	New	HM
MaPLe	90.75	91.84	91.29
+ Ours	90.83	92.00	91.41

(h) FGVCAircraft.

	Base	New	HM
MaPLe	38.34	34.59	36.37
+ Ours	38.00	36.01	36.98

(i) SUN397.

	Base	New	HM
MaPLe	80.92	78.21	79.54
+ Ours	81.04	79.08	80.05

(j) DTD.

	Base	New	HM
MaPLe	79.86	59.54	68.28
+ Ours	79.90	62.28	70.00

(k) EuroSAT.

	Base	New	HM
MaPLe	92.66	68.39	78.70
+ Ours	93.57	71.98	81.37

(l) UCF101.

	Base	New	HM
MaPLe	83.66	78.04	80.75
+ Ours	83.51	79.07	81.23

compared to the baseline. For ChestX, t-SNE embeddings are omitted: the grayscale chest X-rays exhibit minuscule lesion regions that yield highly overlapping clusters, rendering any inter-method distinctions imperceptible.

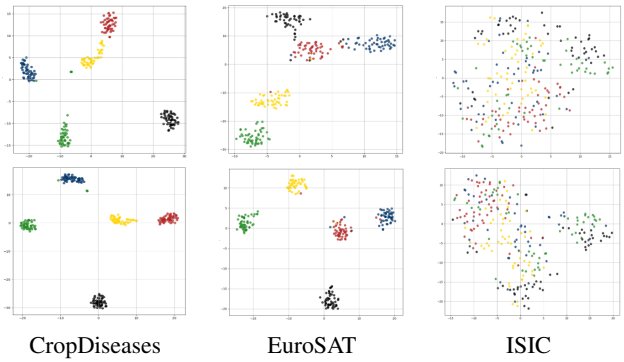

Figure 15. t-SNE visualization of feature distributions on different datasets. The first row shows results from CLIP-LoRA. The second row shows results from our proposed method. Different colors denote different classes.

10. Visualization

10.1. Improved Focus on Relevant Semantics

Figure 16 and Figure 17 respectively present the attention maps and Grad-CAM [30] heatmaps produced by the baseline model and our proposed method across four datasets. This demonstrates that our method enhances the model's ability to focus on critical features in cross-domain few-shot learning tasks and simultaneously improves local alignment between local visual patches and textual semantics.

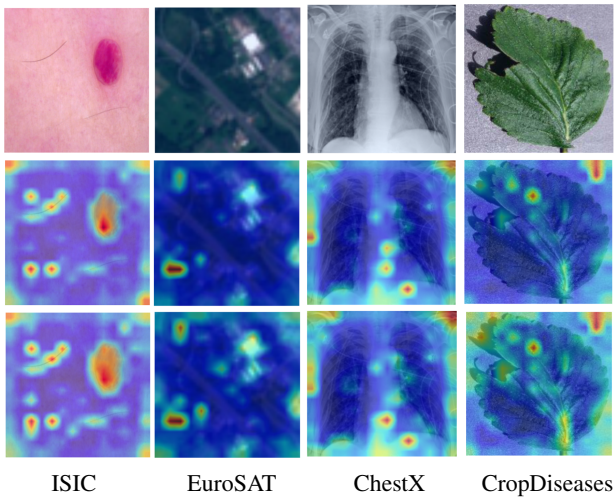

Figure 16. Row 2 and Row 3 display the attention maps of the baseline and CC-CDFSL, respectively, showing that the latter localizes critical regions more precisely.

11. Comparison with SOTA CDFSL Methods

We also evaluate our CC-CDFSL framework in comparison with the most competitive state-of-the-art (SOTA) CDFSL methods, covering a range of settings such as different backbones, the use of source datasets, and whether fine-tuning is performed on the target domain (FT). MEM-FS [34], StyleAdv [9], FLoR [52], DAMIM [26], AttnTemp [53], CD-CLS [54], PMF [16], IM-DCL [37], StepSPT [38], and SeGD-VPT [51] are introduced as competitors. Tabs. 9 and 10 present that our method achieves new SOTA average accuracies of 67.90% and 58.85% for the 5-way 5-shot

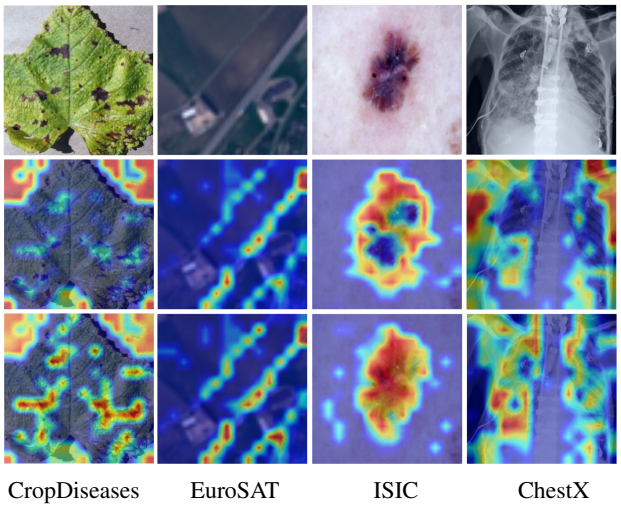

Figure 17. The heatmap for CLIP-LoRA (the second row) and our CC-CDFSL (the third row) in four target domains.

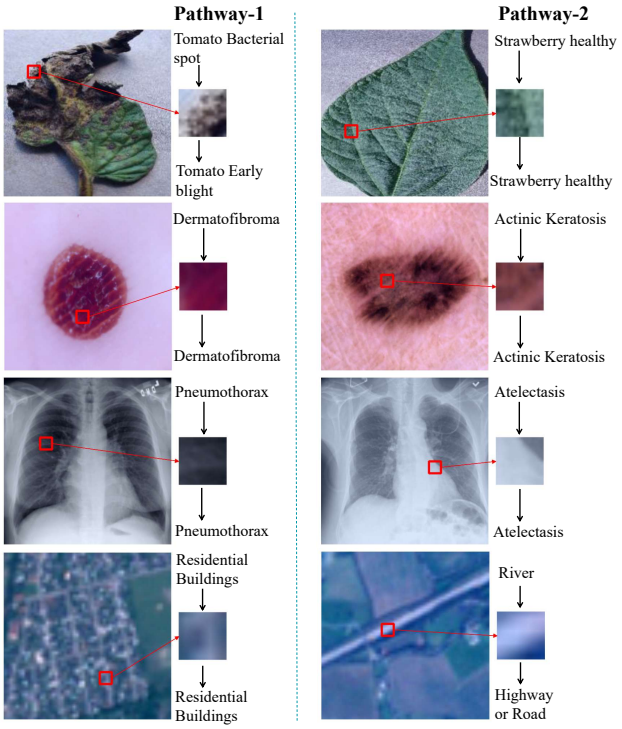

Figure 18. Illustration of the Text-to-Image-patch-to-Text (T-I-T) cycle consistency pathway. Each specific class text is first used to locate the most semantically relevant local patch within the image. The selected image patch is then mapped back to the text space in an attempt to reconstruct the original class label. And the patches in the figure are resized to a resolution of 64×64 for better presentation.

and 5-way 1-shot classification tasks, respectively. Notably, IM-DCL [37] excels on ChestX, which we attribute to its ResNet-based backbone being better suited for capturing local features prevalent in medical images.

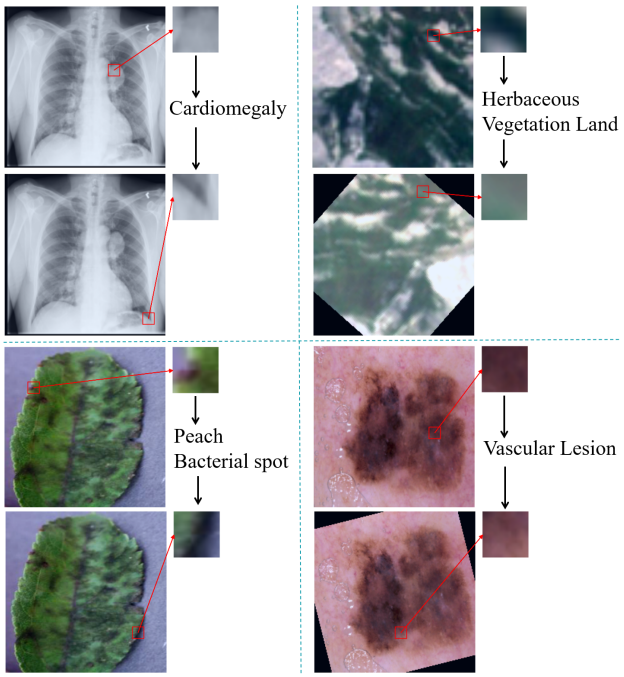

Figure 19. Additional visualizations of the Image-to-Text-to-Image (I-T-I) cycle consistency pathway. Each pathway shows a selected patch within the initial image, the most similar text derived from this patch, and a semantically re-focused patch within its corresponding image in the augmented image space.

12. Interpretability

T-I-T cycle pathway Figure 18 presents additional visualizations of the Text-to-Image-to-Text (T-I-T) cycle consistency pathway. The pathways demonstrate the model's ability to establish fine-grained semantic connections between textual descriptions and visual regions. Even when the reconstructed text does not exactly match the original text (e.g., "River" vs. "Highway or Road" of the EuroSAT for Pathway-2), the inconsistency provides valuable insights into the model's understanding and reasoning at a local level. For instance, both "River" and "Highway or Road" categories represent linear structures (elongated shapes) in satellite imagery. Note that while the Semantic Anchor (SA) Module's augmentation phase expands the corpus for the first hop of the T-I-T cycle, for display convenience, only the original image is used for retrieval in these visualizations.

I-T-I cycle pathway As explicitly mentioned in the main paper, Figure 19 provides more detailed visualizations of the Image-to-Text-to-Image (I-T-I) cycle consistency pathway: the model (1) extracts semantics from an anchor patch (red box), (2) maps it to the most similar text label, and (3) retrieves a semantically matching patch only in the augmented view (red box). It highlights the model's ability to

Table 9. Comparison with state-of-the-art CDFSL works by the 5-way 5-shot classification.

Method	Backbone	Mark	Source	Target	CropDiseases	EuroSAT	ISIC	ChestX	Ave.
MEM-FS	ViT/DINO	TIP-23			93.74	86.49	47.38	26.67	63.57
StyleAdv	ViT/DINO	CVPR-23	✓	-	94.85	88.57	47.73	26.97	64.53
FLoR	ViT/DINO	CVPR-24	✓	-	95.28	90.41	49.52	27.28	65.48
DAMIM	ViT/DINO	AAAI-25	✓	-	95.52	89.50	50.76	27.28	65.77
AttnTemp	ViT/DINO	NeurIPS-24	✓	-	95.53	90.13	53.09	27.72	66.62
CD-CLS	ViT/DINO	NeurIPS-24	✓	✓	96.27	91.53	54.69	27.66	67.54
PMF	ViT/DINO	CVPR-22	✓	✓	92.96	85.98	50.12	27.27	64.08
StyleAdv-FT	ViT/DINO	CVPR-23	✓	✓	95.99	90.12	51.23	26.97	66.08
FLoR-FT	ViT/DINO	CVPR-24	✓	✓	96.47	90.75	53.06	27.02	66.83
DAMIM-FT	ViT/DINO	AAAI-25	✓	✓	96.34	91.18	54.86	27.82	67.78
IM-DCL	RN10	TIP-24	-	✓	95.73	89.47	52.74	**28.93**	66.72
StepSPT	ViT/CLIP	TPAMI-25	-	✓	96.01	89.40	52.12	26.36	65.97
SeGD-VPT	ViT/CLIP	MM-24	-	✓	96.93	93.81	53.10	23.20	66.76
CLIP-LoRA	ViT/CLIP	CVPR-24	-	✓	96.20	92.63	50.68	24.44	65.99
CLIP-LoRA + **Ours**	ViT/CLIP	Ours	-	✓	**97.08**	**94.35**	**54.72**	25.47	**67.90**
Δ	-	-	-	-	**+0.88**	**+1.72**	**+4.04**	**+1.03**	**+1.91**

Table 10. Comparison with state-of-the-art CDFSL works by the 5-way 1-shot classification.

Method	Backbone	Mark	Source	Target	CropDiseases	EuroSAT	ISIC	ChestX	Ave.
MEM-FS	ViT/DINO	TIP-23			81.11	68.11	32.97	22.76	51.24
StyleAdv	ViT/DINO	CVPR-23	✓	-	81.22	72.15	33.05	22.92	52.34
FLoR	ViT/DINO	CVPR-24	✓	-	81.81	72.39	34.20	22.78	52.80
DAMIM	ViT/DINO	AAAI-25	✓	-	82.34	72.87	34.66	22.97	53.21
AttnTemp	ViT/DINO	NeurIPS-24	✓	-	84.02	74.35	34.92	23.19	54.12
CD-CLS	ViT/DINO	NeurIPS-24	✓	✓	84.53	74.97	35.56	23.39	54.62
PMF	ViT/DINO	CVPR-22	✓	✓	80.79	70.74	30.36	21.73	50.91
StyleAdv-FT	ViT/DINO	CVPR-23	✓	✓	84.11	74.93	33.99	22.92	53.99
FLoR-FT	ViT/DINO	CVPR-24	✓	✓	83.55	73.09	35.49	23.26	53.85
DAMIM-FT	ViT/DINO	AAAI-25	✓	✓	83.90	73.61	36.35	23.38	54.31
IM-DCL	RN10	TIP-24	-	✓	84.37	77.14	38.13	**23.98**	55.91
StepSPT	ViT/CLIP	TPAMI-25	-	✓	84.84	70.01	32.97	22.84	52.68
CLIP-LoRA	ViT/CLIP	CVPR-24	-	✓	85.11	81.49	35.23	21.73	55.89
CLIP-LoRA +**Ours**	ViT/CLIP	Ours	-	✓	**88.91**	**86.07**	**38.13**	22.21	**58.83**
Δ	-	-	-	-	**+3.80**	**+4.58**	**+2.90**	**+0.48**	**+2.94**

maintain semantic consistency across different visual transformations and through the textual modality.

13. Generalization beyond CLIP

We focus on the Source-Free Cross-Domain Few-Shot Learning (SF-CDFSL) benchmark established by StepSPT [38] and SeGD-VPT [51], where only a pre-trained model and scarce target-domain shots are available; CLIP is widely adopted as the default backbone in this setting for fair comparison. To verify our findings and method are generalizable to other baselines, we take SigLIP2 [32] and PE-Core [1] as two vision-language baselines with strengthened fine-grained representations. Specifically, SigLIP2 integrates Location-aware Captioners (LocCa) and vision-only self-supervised learning (including SILC and TIPS) to enhance dense prediction and localization capabilities. PE-Core aligns and tunes intermediate-layer features to capture fine-grained spatial representations. As presented in Tab. 11, our method achieves consistent performance gains across diverse backbones.

Table 11. Accuracy on different backbones in 1-shot.

Method	Backbone	ISIC	ChestX	EuroSAT	CropDiseases	Average
CLIP-LoRA	RN50/CLIP	32.01	21.76	57.79	65.24	44.20
+ **OURS**	RN50/CLIP	**35.21**	**22.75**	**59.23**	**72.85**	**47.51**
SigLIP2-LoRA	ViT/SigLip2	26.48	20.53	63.05	81.84	47.98
+ **OURS**	ViT/SigLip2	**29.53**	**22.00**	**68.12**	**83.39**	**50.76**
PE-Core-LoRA	ViT/PE-Core	38.05	22.45	82.16	89.01	57.92
+ **OURS**	ViT/PE-Core	**40.72**	**22.67**	**83.92**	**90.48**	**59.45**

Moreover, we also reproduce the alignment experiments (Fig. 2 in the paper) in Tab. 12. We can see that although these two methods strengthen the fine-grained representations, the phenomenon of degraded local alignment still exists under extreme domain shifts. Quantitatively, Tab. 12 shows that our method consistently improves local alignment scores by **+26.3%** for SigLIP2 and **+18.6%** for PE-Core, despite different ways to pretrain the VLM.

In summary, these experiments verify that our findings

Table 12. Feature alignments of other baselines.

Method	Feature type	ISIC	ChestX	EuroSAT	CropDiseases	Average
SigLIP2-LoRA	global	0.1511	0.1380	0.1226	0.1691	0.14520
SigLIP2-LoRA	local	0.0330	0.0310	0.0126	0.0383	0.02872
+ OURS	local	0.0421	0.0363	0.0181	0.0486	0.03627
PE-Core-LoRA	global	0.2795	0.2493	0.2459	0.2810	0.26392
PE-Core-LoRA	local	0.0191	0.0096	0.0264	0.0088	0.01597
+ OURS	local	0.0238	0.0111	0.0286	0.0123	0.01895

and designs are generalizable to other baselines.

14. Hyperparameters in Eq. 16

Table 13 lists the cycle-consistency weights λ_1 (T-I-T) and λ_2 (I-T- I) for each target dataset, determined by grid search on the validation split. All experiments use $k = 10$ for selecting anchor patches.

Table 13. Hyperparameters λ_1 and λ_2 for each dataset.

	ChestX	ISIC2018	EuroSAT	CropDiseases
λ_1	3	3	1.5	1
λ_2	0.5	2	0.2	1.5