Addressing Exacerbated Attention Sink for Source-Free Cross-Domain Few-Shot Learning

Supplementary Material

7. Detailed Dataset Description

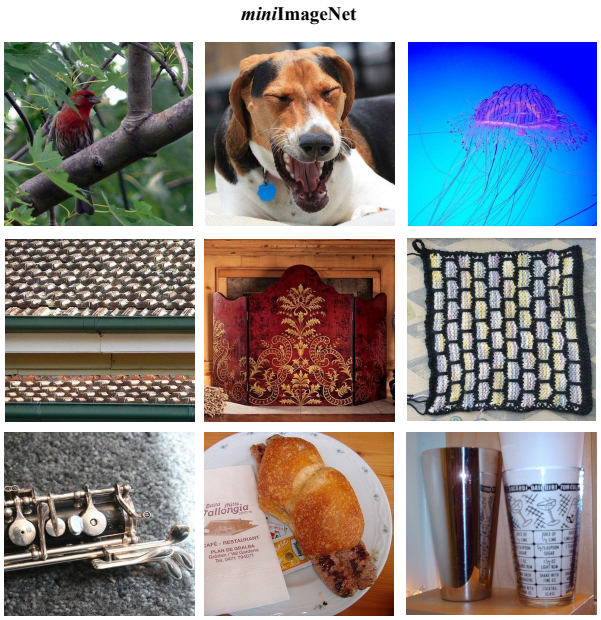


Figure 9. Representative samples from the source-domain *mini*ImageNet dataset.

***mini*ImageNet** [24] is a widely adopted benchmark in meta-learning and few-shot learning, comprising a curated subset of the original ImageNet [6] dataset. The dataset contains 60,000 color images distributed across 100 object categories, with each category consisting of 600 samples of size 84×84 pixels. As illustrated in Fig. 9, *mini*ImageNet encompasses diverse real-world scenes with varying contextual elements, including both human-centric scenarios and natural environments.

CropDiseases [20] provides a specialized agricultural dataset for plant disease recognition, containing 43,456 high-resolution images across 38 disease categories. The dataset exhibits huge domain shift from natural images, featuring detailed close-ups of infected and healthy plant specimens with high intra-class similarity.

EuroSAT [11] offers a remote sensing benchmark for land use classification, comprising 27,000 satellite images categorized into 10 distinct land cover types. The aerial perspective and absence of conventional photographic distortions create substantial domain gap from natural image datasets.

ISIC2018 [3] constitutes a medical imaging dataset for dermatological analysis, containing 10,015 dermoscopic images across 7 skin lesion categories. The clinical nature and specialized visual patterns represent huge domain shift from conventional natural images.

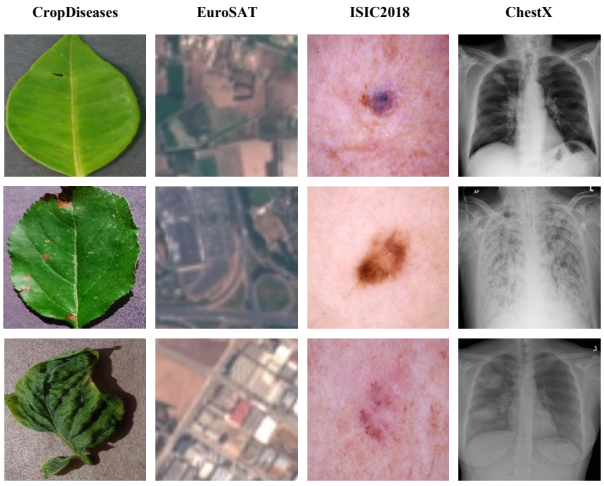


Figure 10. Sample images from the four target-domain benchmarks: CropDiseases (agricultural), EuroSAT (remote sensing), ISIC2018 (dermatological), and ChestX (medical imaging).

ChestX [25] provides a medical radiology dataset of 25,847 chest X-ray images distributed across 7 thoracic conditions. This dataset exhibits the most substantial domain gap due to its monochromatic medical imaging modality, anatomical focus, and absence of natural scene characteristics.

As depicted in Fig. 10, these four target domains: agricultural, remote sensing, dermatological, and medical radiology, which collectively represent challenging cross-domain scenarios with progressively increasing domain shifts from the source domain.

8. Detailed Descriptions of the CKA

Following established practices in domain similarity measurement [18, 21], we employ Centered Kernel Alignment (CKA) to quantitatively assess the similarity between feature representations across different domains. CKA is a robust statistical method specifically designed to compare high-dimensional representations learned by neural networks, with particular strength in analyzing cross-domain feature relationships.

The CKA methodology operates through the following computational process. Given two sets of feature representations $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times d}$ extracted from different domains, we first compute their Gram matrices:

$$K = XX^\top, \quad L = YY^\top \tag{8}$$

These matrices capture the inner product relationships between all pairs of samples within their respective feature spaces. To eliminate the influence of data means and ensure proper alignment, we center the Gram matrices using:

$$K_c = HKH \tag{9}$$

$$L_c = HLH \tag{10}$$

where $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ denotes the centering matrix, I_n is the identity matrix, and $\mathbf{1}_n$ represents a vector of ones. The final CKA similarity is then computed as:

$$\text{CKA}(X, Y) = \frac{\text{vec}(K_c) \cdot \text{vec}(L_c)}{\|\text{vec}(K_c)\|\|\text{vec}(L_c)\|} = \frac{\text{Tr}(K_c L_c)}{\sqrt{\text{Tr}(K_c^2)\text{Tr}(L_c^2)}} \tag{11}$$

This normalized metric quantifies the similarity between the relational structures encoded in the two feature sets, with values ranging from 0 (completely dissimilar) to 1 (identical relationships).

In our work, we leverage CKA to measure domain distance between source and target datasets following [5]. Given a pre-trained backbone network, we extract features from image batches sampled from different domains and compute CKA similarity after aligning the channel dimensions. The interpretation follows established conventions [17]: higher CKA values indicate greater domain similarity and consequently less domain-specific information in the representations, while lower values suggest stronger domain shift and more domain-characteristic features. This approach provides crucial insights into the model's generalization capabilities and domain adaptation characteristics across diverse data distributions.

9. More Experiments

9.1. Norm distribution of different sum numbers from layer 0 to layer 11

We provide comprehensive layer-wise analyses of token norm distributions across all transformer layers (0 to 11) in Fig.11. The complete visualization across all layers offers deeper insights into the evolution of semantic awareness and attention patterns throughout the network architecture.

As systematically illustrated in the Fig.11, several key patterns emerge across the layer hierarchy:

(1) **Early Layers (0-4)**: Both source and target domains exhibit similar norm distributions regardless of sum scores, confirming that shallow layers primarily process low-level features with minimal semantic discrimination. The uniform distribution across all token types demonstrates the model's initial lack of class-specific awareness in these foundational layers.

(2) **Middle Layers (5-7)**: A gradual divergence begins to emerge between source and target domains. In the source domain, we observe the initial development of differential norm patterns, with "Sum=1" tokens starting to receive slightly elevated attention. Conversely, target domains maintain relatively flat distributions, indicating stalled development of discriminative patterns due to domain shift.

(3) **Deep Layers (8-11)**: The critical divergence becomes most pronounced in these final layers. Source domain models exhibit strong specialization, with "Sum=1" discriminative tokens receiving substantially higher norms than other token types. This pattern reflects successful development of class-specific attention mechanisms. In contrast, target domain models show two distinct trajectories: pre-trained models maintain weak discrimination, while fine-tuned models develop inverted patterns with "Sum=5" sink tokens dominating the norm distribution.

The complete layer progression provides compelling evidence for our core thesis: the attention sink exacerbation stems from fine-tuning's impact on deep layer transformations rather than cross-domain transfer alone. While domain shift prevents the natural emergence of discriminative patterns, it is the fine-tuning process that actively drives the model toward non-discriminative sink tokens as an adaptation strategy.

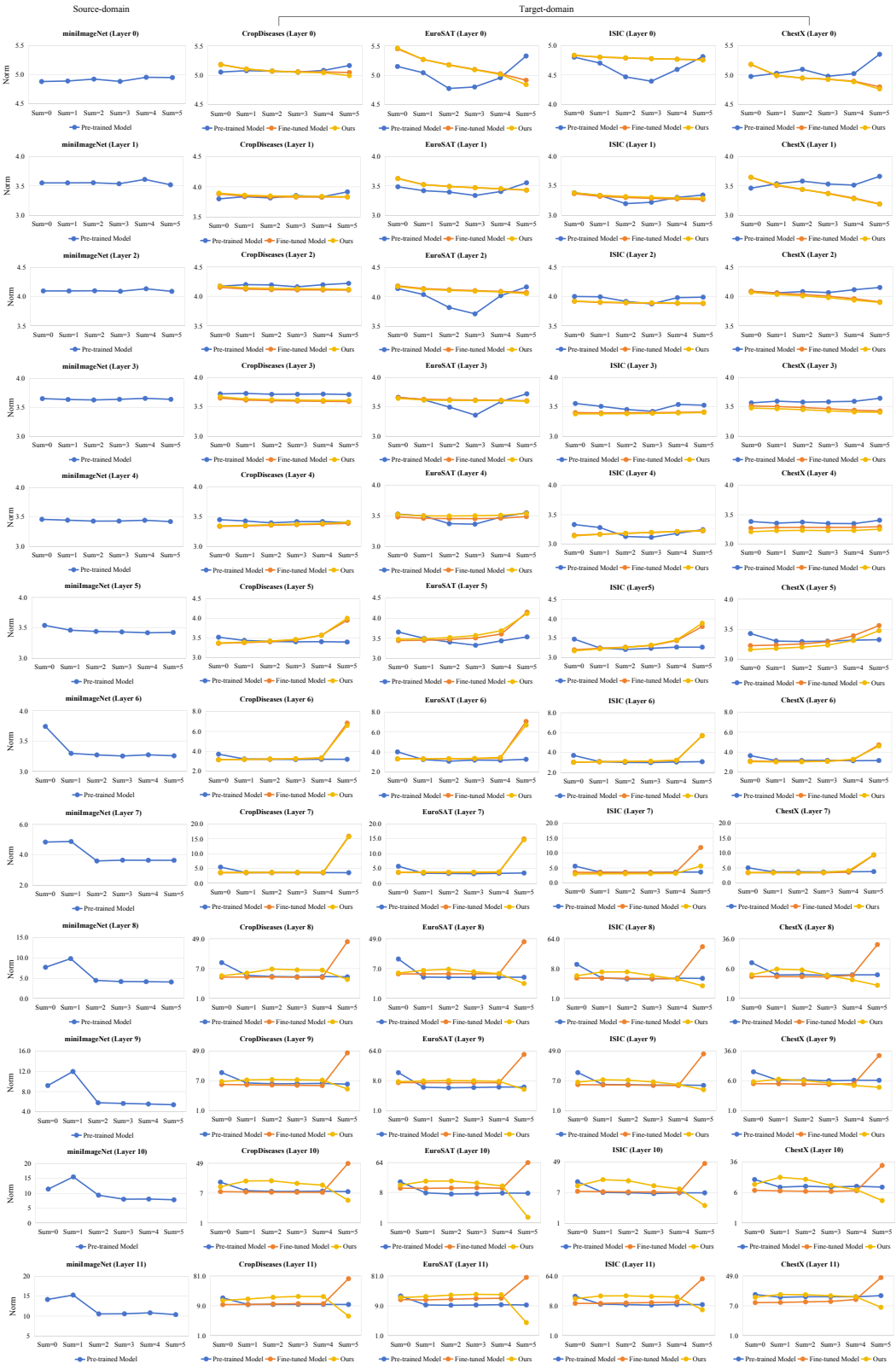
These comprehensive results across all layers and datasets reinforce our interpretation that standard fine-tuning in CDFSL scenarios creates a problematic shortcut, where models prioritize easily alignable but non-discriminative patterns over semantically meaningful features, ultimately limiting their generalization capability in target domains.

Table 3. TIR with more backbones.

Method (Shot)	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
ViT/SigLIP2 (1-shot)	80.98	66.52	29.04	21.34	49.47
+Ours	**84.82**	**69.87**	**32.20**	**21.53**	**52.11**
ViT/PE-Core (1-shot)	87.16	78.43	39.39	21.93	56.73
+Ours	**88.84**	**80.17**	**40.31**	**22.23**	**57.89**
ViT/SigLIP2 (5-shot)	95.43	85.17	44.26	23.68	62.14
+Ours	**97.05**	**88.83**	**51.10**	**24.61**	**65.40**
ViT/PE-Core (5-shot)	96.97	90.46	55.22	25.52	67.04
+Ours	**97.21**	**92.60**	**57.34**	**25.89**	**68.26**

9.2. Extending to more backbones

We evaluate our method on two representative backbones, SigLIP2 [23] and PE-Core [2], with results reported in Tab. 3. Our approach consistently achieves improvements over the baselines in both the 1-shot and 5-shot settings, demon-



Norms of Pre-trained Model in source-domain dataset (miniImageNet) and Norms of three models in target-domain datasets (CropDiseases, EuroSAT, ISIC, ChestX) in layer 0-11

Figure 11. At shallow layers, both source and target domain models show similar attention distributions with weak semantic awareness. At deep layers, the source domain model develops specialized attention to discriminative tokens ("Sum=1"), while the target domain fine-tuned model shifts toward non-discriminative tokens ("Sum=5") with higher norms, demonstrating that fine-tuning, instead of merely cross-domain transfer, serves as the primary driver of attention sink exacerbation.

Table 4. TIR with more adaptation strategies.

Method (Shot)	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
CLIP + CoOp	91.85	83.27	43.31	22.67	60.28
+Ours	**92.73**	**85.33**	**44.49**	**24.44**	**61.75**
CLIP + LoRA(Text)	92.74	82.98	42.97	22.84	60.38
+Ours	**93.19**	**83.26**	**47.35**	**23.33**	**61.78**
CLIP + MaPLe	96.20	90.00	50.47	24.11	65.20
+Ours	**96.80**	**93.35**	**54.04**	**25.75**	**67.49**
CLIP + LoRA(Vision)	96.21	92.52	51.10	24.13	65.99
+Ours	**97.42**	**93.49**	**56.73**	**26.12**	**68.44**

Table 5. Comparisons with three types of attention sink methods.

Method	Mark	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
Baseline	-	96.21	92.52	51.10	24.13	65.99
SPARC	ICML25	70.60	62.61	40.58	22.70	49.12
Registers	NeurIPS25	96.54	90.67	50.53	23.56	65.33
PAINT	CVPRW25	96.88	91.65	50.58	23.62	65.68
TIR	Ours	**97.42**	**93.49**	**56.73**	**26.12**	**68.44**

strating its effectiveness across different architectures and few-shot scenarios.

9.3. Experiments with other finetuning strategies

We implement TIR with more fine-tuning strategies in Tab. 4, and observe consistent improvements across all settings. Specifically, when applied to CoOp [45], LoRA [12] (Text), MaPLe [16], and LoRA [12] (Vision), our method yields higher accuracy on each individual dataset. Notably, the gains are particularly pronounced on the more challenging ISIC2018 and ChestX datasets, and the overall average improvement reaches up to 2.45% when combined with LoRA (Vision). These results demonstrate that TIR generalizes effectively across diverse adaptation strategies, consistently boosting few-shot performance.

9.4. Comparison with other attention sink works

We compare our method with existing approaches designed to mitigate attention sink phenomena, including SPARC [14], Registers [4], and PAINT [1]. As shown in Tab. 5, these methods either underperform or yield only marginal improvements over the baseline in the source-free cross-domain few-shot learning setting. Specifically, SPARC causes a substantial performance drop across all datasets, achieving an average accuracy of only 49.12%. Registers and PAINT obtain averages of 65.33% and 65.68%, respectively. Both are slightly below the baseline's 65.99%. In contrast, our method consistently improves upon the baseline across every dataset, attaining the highest average accuracy of 68.44%. These results demonstrate the effectiveness of our approach in addressing attention sink challenges under domain shift and few-shot constraints.

9.5. Other design choices

We present a comprehensive ablation study examining various design choices for our token recalibration strategy. The experimental results, summarized in the Tab.6, provide

Table 6. Comprehensive ablation study of different design choices for token recalibration on the 5-way 5-shot task.

Method	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
Baseline	96.21	92.52	51.10	24.13	65.99
Only "Sum=5" weight=0	97.21	93.26	55.00	25.15	67.66
Only "Sum=1" weight=2	97.24	92.99	53.72	24.98	67.23
Ours	97.42	**93.49**	**56.73**	**26.12**	**68.44**
"Sum=5" weight=0 and "Sum=1" weight=2	**97.43**	93.40	56.72	25.92	68.37
Only "Sum=5" weight=0.5	96.22	92.54	52.42	24.32	66.38
Only "Sum=5" weight=0.1	96.40	92.60	52.75	24.61	66.59
Only "Sum=1" weight=1.5	97.18	92.96	53.47	24.87	67.12
Only "Sum=1" weight=3	97.22	92.83	53.44	24.81	67.08
Only "Sum=1" weight=4	97.15	92.79	53.43	24.96	67.08
Only "Sum=1" weight=5	97.19	92.82	53.37	24.65	67.01

valuable insights into the optimal configuration for addressing attention sink in SF-CDFSL scenarios.

(1) **Critical Role of Sum=5 Suppression**: The experimental results demonstrate that completely suppressing Sum=5 tokens (setting their weights to 0) yields the most significant performance improvements. Alternative suppression strengths (weights of 0.5 or 0.1) prove substantially less effective, indicating that strong suppression of these non-discriminative sink tokens is essential for mitigating attention sink exacerbation.

(2) **Robustness of Sum=1 Enhancement**: Interestingly, the enhancement of Sum=1 discriminative tokens shows remarkable robustness to specific weight values. Our experiments with enhancement weights of 1.5, 3, 4, and 5 reveal that any value greater than 1 produces comparable performance gains. This insensitivity to the exact enhancement magnitude suggests that the crucial factor is simply providing positive reinforcement to class-specific tokens rather than fine-tuned weight optimization.

(3) **Effectiveness of Minimal Intervention**: Perhaps most notably, the combination of solely suppressing Sum=5 tokens while enhancing Sum=1 tokens achieves performance nearly equivalent to our complete method (which additionally handles intermediate Sum=2,3,4 tokens). This finding indicates that the core attention sink problem in SF-CDFSL primarily stems from the extreme cases of completely non-discriminative and highly discriminative tokens.

These insights lead to a simplified and generalizable approach for SF-CDFSL: in K-way N-shot settings, setting the weight of Sum=K tokens to 0 while assigning Sum=1 tokens any weight greater than 1 effectively addresses the attention sink exacerbation problem. This simplified strategy maintains the performance benefits of our complete method while offering greater practicality and easier implementation for various few-shot learning configurations.

The robustness of this approach across different enhancement weights and its effectiveness with minimal intervention make it particularly suitable for real-world SF-CDFSL applications where architectural simplicity and parameter insensitivity are valuable attributes.