Appendix

Contents

A. Detailed Related Work

A.1. Source-Free Cross-domain Few shot Learning

Cross-Domain Few-Shot Learning (CDFSL) aims to train a model on a source domain that can generalize effectively to a target domain with limited examples. Existing methods are typically categorized into two types: meta-learning-based approaches [12, 15, 21, 48, 61] and transfer learning-based approaches [15, 17, 33, 67, 73–75]. Source-Free Cross-Domain Few-Shot Learning (SF-CDFSL) introduces a stronger constraint by making source domain data inaccessible. Current SF-CDFSL methods [54, 58, 71] primarily rely on large models, such as CLIP [40], leveraging their prior knowledge for classification in the target domain. However, these approaches fail to account for the misalignment between modalities when transferring CLIP to cross-domain settings. Moreover, the influence of visual learning on CLIP-based SF-CDFSL tasks remains underexplored.

A.2. Parameter-Efficient Fine-Tuning

Efficiently applying Vision-Language Models (VLMs) to downstream tasks is a key research area. A common strategy is parameter-efficient fine-tuning (PEFT), which uses only a few samples from the target task. PEFT adjusts a small number of the VLM's parameters, allowing the model to adapt to various applications without modifying all pre-trained parameters. PEFT methods can be grouped into three main types: prompt learning, adapters, and LoRA (and its variants). Prompt learning transforms fixed templates into learnable parameters, such as CoOp [69] , CoCoOp [68], MaPLe [25], PLOT [6], ProGrad [70], PromptSRC [27], KgCoOp [56], PCB [2], DynaPrompt [52], TCP [57] and ATPrompt [32]. Additionally, Customized Ensemble [35] combines outputs from multiple models for improved performance, and PromptKD [31] explores

knowledge distillation in prompt learning. Adapter-based methods add trainable modules to the original frozen architecture, making fine-tuning easier, such as CLIP-Adapter [14], Tip-Adapter [64], LP++ [22], AMU-Tuning [44], LatHAdapter [66], MMA [55] and LDC [30]. Low-Rank Adaptation (LoRA) [19, 60] fine-tunes the model by adding learnable low-rank matrices while keeping the original parameters fixed. The new weights can be merged with the original ones, and LoRA does not add extra inference time. Various studies have extended LoRA by adapting the rank for each matrix [47, 62], improving its performance [5, 28, 72], or reducing memory usage through quantization [10, 41]

A.3. Modality Gap and Misalignment

[34] was the first to identify the existence of a modality gap in multimodal models. Some reserch observed that the performance of multimodal models significantly declines when facing significant domain shifts [24, 42]. [23] highlighted the cross-modal bias between semantic-guided samples and nonsemantic-guided samples. [16] emphasized the pairwise misalignment of multimodal uncertainties, addressing the one-to-many alignment issue in multimodal video-text retrieval tasks. [45] pointed out that ProtoNet exhibits a gap between prototypes and instances in cross-domain scenarios. [49] attributed generalization errors to the model learning non-aligned features between the source domain and the test data. These existing works identified the issue of modality misalignment in cross-domain scenarios and considered fine-tuning an effective method for realignment. However, our work demonstrates that fine-tuning alone is insufficient for effective realignment in cross-domain scenarios, especially with large domain gaps in the CDFSL task, such as general domains vs. medical domains. We analyze this issue from the perspective of visual learning, which acts as a shortcut during fine-tuning, and address it in this paper.

B. Proof of Theorem 4.1

The loss in CLIP is designed to align image-text pairs while contrasting mismatched pairs. Below is the step-by-step derivation of the gradient for the visual feature f_i. For a batch of N image-text pairs, the loss for the i-th image is:

$$\mathcal{L}_i = -\log \frac{e^{f_i \cdot t_i/\tau}}{\sum_{k=1}^{N} e^{f_i \cdot t_k/\tau}}, \tag{1}$$

where: f_i: L2-normalized visual feature of the i-th image ($\|f_i\| = 1$). t_k: L2-normalized text feature of the k-th text ($\|t_k\| = 1$). τ: Temperature coefficient (e.g., $\tau = 0.01$).

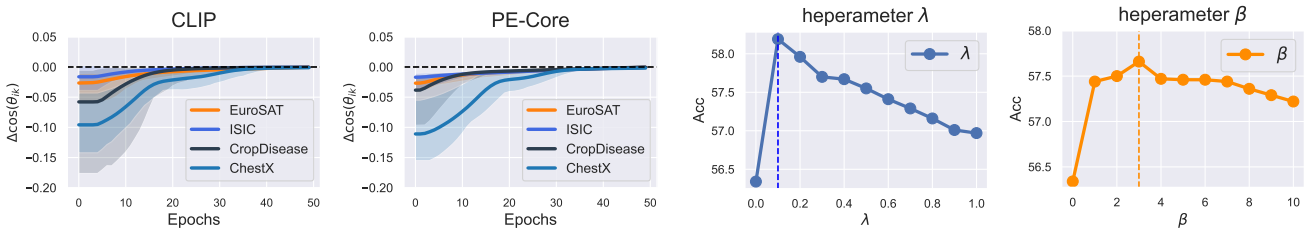
Expand $\mathcal{L}_i$:

$$\begin{aligned}\mathcal{L}_i &= \underbrace{-\log\left(e^{f_i \cdot t_i/\tau}\right)}_{\text{postive}} + \underbrace{\log\left(\sum_{k=1}^{N} e^{f_i \cdot t_k/\tau}\right)}_{\text{negative}} \\ &= -\frac{f_i \cdot t_i}{\tau} + \log\left(\sum_{k=1}^{N} e^{f_i \cdot t_k/\tau}\right).\end{aligned} \tag{2}$$

The gradient $\nabla_{f_i}\mathcal{L}_i$ has two terms from the loss components:

$$\begin{aligned}\nabla_{f_i}\mathcal{L}_i &= -\nabla_{f_i}\left(\frac{f_i \cdot t_i}{\tau}\right) + \nabla_{f_i}\left(\log\left(\sum_{k=1}^{N} e^{f_i \cdot t_k/\tau}\right)\right) \\ &= -\frac{t_i}{\tau} + \frac{1}{\sum_{m=1}^{N} e^{f_i \cdot t_m/\tau}} \cdot \sum_{k=1}^{N} \nabla_{f_i}\left(e^{f_i \cdot t_k/\tau}\right) \\ &= -\frac{t_i}{\tau} + \frac{\sum_{k=1}^{N} e^{f_i \cdot t_k/\tau} \cdot \frac{t_k}{\tau}}{\sum_{m=1}^{N} e^{f_i \cdot t_m/\tau}} \\ &= -\frac{t_i}{\tau} + \frac{1}{\tau}\sum_{k=1}^{N} p_{ik} t_k\end{aligned} \tag{3}$$

where p_{ik} is the softmax probability: $p_{ik} = \frac{e^{f_i \cdot t_k/\tau}}{\sum_{m=1}^{N} e^{f_i \cdot t_m/\tau}}$. Let $\mathbf{f}_i^{\text{new}}$ and $\mathbf{f}_k^{\text{new}}$ represent the visual features after parameter updates. The term $\mathbf{f}_i^{\text{new}} \cdot \mathbf{f}_k^{\text{new}}$ denotes the similarity between samples i and k after the parameter update, then:



(a) $\Delta\cos(\theta_{ik})$ in fine-tuning (b) Hyperparameter study

Figure 1. (a) When sample i and sample k belong to different classes, $\Delta\cos(\theta_{ik})$ in 5-way 1-shot fine-tuning is always less than 0, which means that the visual features of samples i and k from different classes will become increasingly dissimilar. (b) Hyperparameter study.

$$\mathbf{f}_i^{\text{new}} = \mathbf{f}_i + \eta\frac{1}{\tau}\left(\mathbf{t}_i - \sum_{j=1}^{N} p_{ij}\mathbf{t}_j\right), \mathbf{f}_k^{\text{new}} = \mathbf{f}_k + \eta\frac{1}{\tau}\left(\mathbf{t}_k - \sum_{j=1}^{N} p_{kj}\mathbf{t}_j\right) \tag{4}$$

$$\begin{aligned}\mathbf{f}_i^{\text{new}} \cdot \mathbf{f}_k^{\text{new}} &= (\mathbf{f}_i + \eta\frac{1}{\tau}\left(\mathbf{t}_i - \sum_{j=1}^{N} p_{ij}\mathbf{t}_j\right)) \cdot (\mathbf{f}_k + \eta\frac{1}{\tau}\left(\mathbf{t}_k - \sum_{j=1}^{N} p_{kj}\mathbf{t}_j\right)) \\ &= \mathbf{f}_i \cdot \mathbf{f}_k + \eta\frac{1}{\tau}\left(\mathbf{f}_i \cdot (\mathbf{t}_k - \sum_{j=1}^{N} p_{kj}\mathbf{t}_j) + \mathbf{f}_k \cdot (\mathbf{t}_i - \sum_{j=1}^{N} p_{ij}\mathbf{t}_j)\right) + O(\eta^2),\end{aligned} \tag{5}$$

where η is the learning rate and τ is the temperature coefficient. Let the difference in cosine similarity between two samples, $\mathbf{f}_i$ and $\mathbf{f}_k$, before and after one step of training is $\Delta\cos(\theta_{ik}) = \mathbf{f}_i^{\text{new}} \cdot \mathbf{f}_k^{\text{new}} - \mathbf{f}_i \cdot \mathbf{f}_k$, there is:

$$\begin{aligned}\Delta\cos(\theta_{ik}) &= \mathbf{f}_i^{\text{new}} \cdot \mathbf{f}_k^{\text{new}} - \mathbf{f}_i \cdot \mathbf{f}_k \\ &= \eta\frac{1}{\tau}\left(\mathbf{f}_i \cdot \mathbf{t}_k - \sum_{j=1}^{N} p_{kj}\mathbf{f}_i \cdot \mathbf{t}_j + \mathbf{f}_k \cdot \mathbf{t}_i - \sum_{j=1}^{N} p_{ij}\mathbf{f}_k \cdot \mathbf{t}_j\right) + O(\eta^2)\end{aligned} \tag{6}$$

When $\mathbf{f}_i$ and $\mathbf{f}_k$ belong to the same class, $\mathbf{t}_i = \mathbf{t}_k$. Considering that the learning objective of the cross-entropy loss function satisfies: $\mathbf{f}_i \cdot \mathbf{t}_i > \mathbf{f}_i \cdot \mathbf{t}_j, \forall j \neq i$. there is:

$$\begin{aligned}\Delta\cos(\theta_{ik}) &= \eta\frac{1}{\tau}\left(\mathbf{f}_i \cdot \mathbf{t}_i - \sum_{j=1}^{N} p_{kj}\mathbf{f}_i \cdot \mathbf{t}_j + \mathbf{f}_k \cdot \mathbf{t}_k - \sum_{j=1}^{N} p_{ij}\mathbf{f}_k \cdot \mathbf{t}_j\right) + O(\eta^2) \\ &> \eta\frac{1}{\tau}\left(\mathbf{f}_i \cdot \mathbf{t}_i - (\sum_{j=1}^{N} p_{kj})\mathbf{f}_i \cdot \mathbf{t}_i + \mathbf{f}_k \cdot \mathbf{t}_k - (\sum_{j=1}^{N} p_{ij})\mathbf{f}_k \cdot \mathbf{t}_k\right) \\ &= \eta\frac{1}{\tau}\left(\mathbf{f}_i \cdot \mathbf{t}_i - 1 \cdot \mathbf{f}_i \cdot \mathbf{t}_i + \mathbf{f}_k \cdot \mathbf{t}_k - 1 \cdot \mathbf{f}_k \cdot \mathbf{t}_k\right) = 0.\end{aligned} \tag{7}$$

That is, the visual features between samples i and k of the same class will become increasingly similar.

When $\mathbf{f}_i$ and $\mathbf{f}_k$ belong to different classes, we analyzed the values of $\Delta\cos(\theta_{ik})$ during the training process across 200 episodes, as shown in Figure 1. As seen, in different VLMs (CLIP [40] and PE-Core [3]), the $\Delta\cos(\theta_{ik})$ is always less than 0., which means that the visual features of samples i and k from different classes will become increasingly dissimilar.

Summary: during fine-tuning, visual features of the same category cluster together while those of different categories are pushed apart, which means visual learning is consistently present in the fine-tuning process.

C. Hyperparameter Study

Our method involves two main hyperparameters: λ, corresponding to the Suppressing Visual Learning (SVL) module, and β, corresponding to the Relationship Alignment (RA) module. Here, we present the 5-way 1-shot performance corresponding

to different values of λ and β on the four CDFSL datasets, as shown in Figure 1b. As observed, the model achieves peak performance when λ is set to 0.1 and β to 3. A smaller value for λ is preferred, as $\mathcal{L}_{ad}$ functions as a perturbation and should be assigned a small weight to prevent it from dominating the training process.

D. Alternative Strategies for the SVL Module

In this section, we discuss several alternative strategies for the SVL modules. The SVL module is designed to perturb the visual learning, thereby guiding the model to focus more on learning cross-modal relationships. Our approach involves randomly sampling support samples to create class-shuffle weights and performing the classification task on the visual branch (as described in Section 5.1 of the main text). To demonstrate the effectiveness of this strategy, we also test three other approaches: **NO**, which does not use the SVL module and serves as the baseline, and $-\boldsymbol{\mathcal{L}_v}$, which use the negative of $\mathcal{L}_v$ as $\mathcal{L}_{ad}$, where $\mathcal{L}_{ad} = -\mathcal{L}_v$ (as defined in Equation 3 of the paper). Additionally, **Noise Proto** generates class-shuffle weights using random initialization (non-true visual features) and then performs the classification task on the visual branch.

Table 1. Alternative strategies for the SVL module.

Disturb Strategy	Cropdisease	EuroSAT	ISIC	ChestX	Avg
No	84.62	81.72	36.40	21.86	56.07
$-\mathcal{L}_v$	82.41	78.84	34.72	21.37	54.33
Noise Proto	84.31	82.42	35.96	21.80	56.12
Ours	**86.41**	**83.80**	**37.63**	**22.38**	**57.55**

Table 2. Alternative strategies for the RA module.

Alignment Strategy	Cropdisease	EuroSAT	ISIC	ChestX	Avg
No	84.62	81.72	36.40	21.86	56.07
only vision	84.87	82.61	36.94	22.11	56.63
only text	84.14	81.36	36.87	21.34	55.92
Ours	**85.92**	**83.21**	**37.42**	**22.17**	**57.17**

The results, shown in Table 1, indicate that the $-\boldsymbol{\mathcal{L}_v}$ strategy actually harms the model's performance. Moreover, the **Noise Proto** strategy brings little or no improvement in performance. This is understandable, as a fundamental requirement for visual learning is that the learning occurs within a valid visual feature distribution. The class-shuffle weights generated by the **Noise Proto** strategy may not be within this valid distribution. In contrast, our approach ensures the simulation of an effective visual feature distribution while simultaneously suppressing discriminative learning, leading to a significant improvement in model performance.

E. Alternative Strategies for the RA Module

In this section, we discuss several alternative strategies for the RA modules. The RA module is designed to replace the discriminative visual feature learning direction in the fine-tuning process, providing a new learning direction that facilitates alignment between modalities. Our approach promotes cross-modal alignment by gradually aligning the internal relationships of visual features with those of textual features (as described in Section 5.2 of the main text). To demonstrate the effectiveness of this strategy, we also test three other approaches: **NO**, which does not use the RA module and serves as the baseline, and **only vision**, which does not incorporate the internal relationships of the text modality and only maintains the relationships between visual features, i.e., $\mathcal{L}_{\text{ra}} = D_{KL}(A^v||A^v)$ (see Equation 10 in the main text). Additionally, **only text** does not use the gradual fusion approach (Equation 9 in the main text) but directly aligns the internal relationships of visual features with those of textual features, i.e., $\mathcal{L}_{\text{ra}} = D_{KL}(A^v||A^t)$.

The results, shown in Table 2, indicate that the **only text** strategy harms the model's performance. In contrast, our method, which gradually fuses the relationships between visual features A_v and textual features A_t to guide the learning of internal visual feature relationships, most effectively promotes alignment between modalities and achieves optimal performance.

F. Division Between Initial Epochs and Later Epochs

During the fine-tuning process, we train the model for 250 epochs, with the first 3/5 epochs used as the initial training phase. In this section, we conduct further experiments and analysis on the initial training phase. We test different proportions for the initial training phase division, as shown in Figure 2. The results indicate that when 150 epochs (the first 3/5) is used as the initial training phase, the model achieves optimal performance across four datasets. It is important to note that in this section, we focus only on using the initial epochs for the initial training phase, as Section 6.5 of the main text

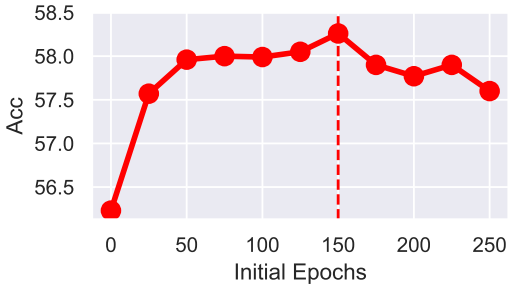


Figure 2. Results for different initial epochs settings.

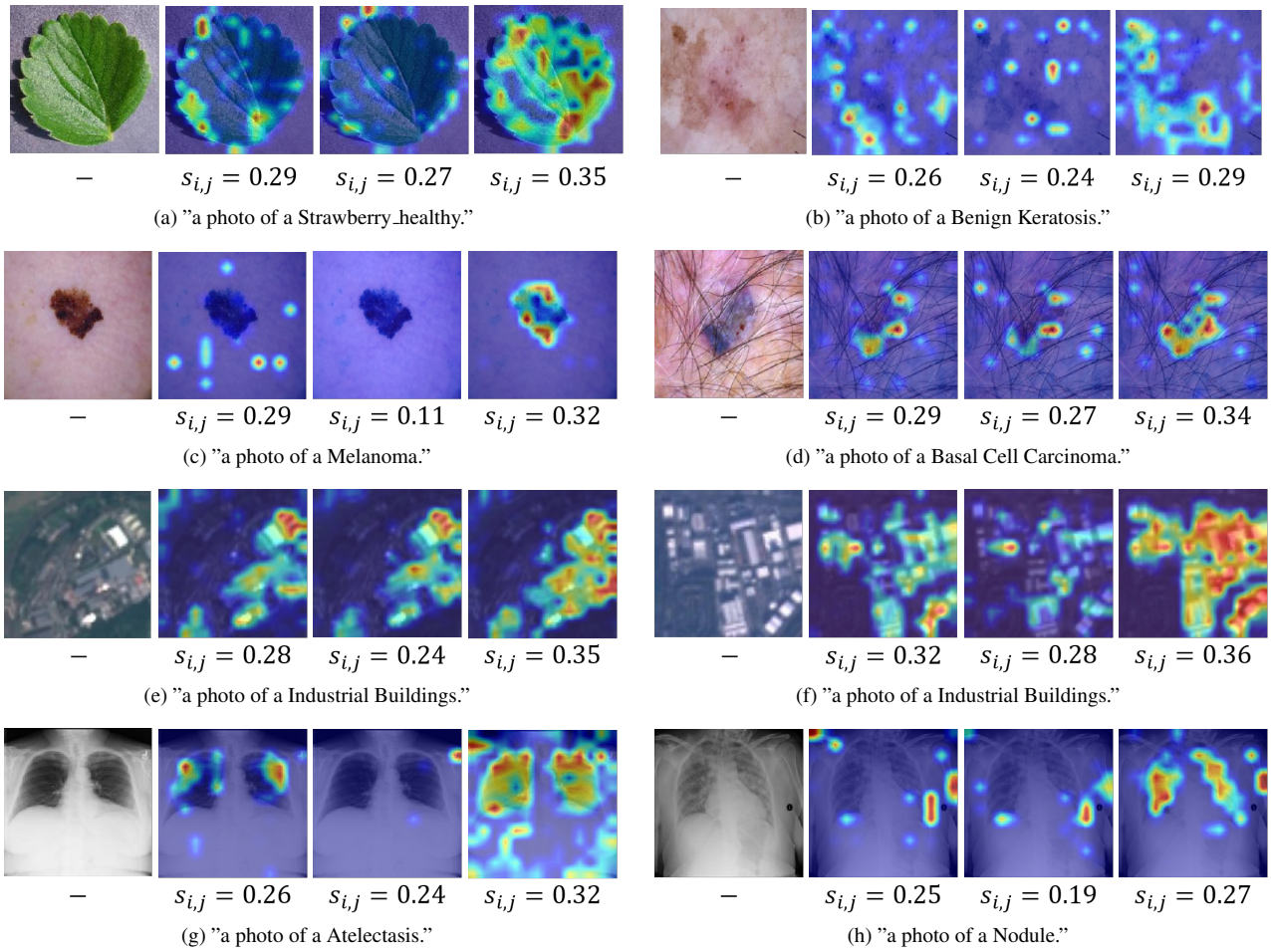


Figure 3. The attention maps of the three models are shown. From left to right: the original image, the Baseline result (trained with $\mathcal{L}_{\text{vlm}}$), the result of the model with enhanced visual learning (trained with $\mathcal{L}_{\text{vlm}} + \mathcal{L}_{\text{v}}$), and the result of our method (trained with $\mathcal{L}_{\text{vlm}} + \mathcal{L}_{\text{ad}}$). Here, $s_{i,j}$ represents the cosine similarity between the image features and the text features. A higher similarity indicates better alignment. It is evident that our model has the most appropriate attention scope and achieves the highest similarity between sample pairs.

has already demonstrated the need to suppress visual learning in the early stages.

G. Better Modality Alignment

Additionally, the modules we proposed effectively facilitate the cross-modal alignment process. Firstly, as shown in Table 1 of the main text, both of our proposed modules improves cross-modal classification performance while reducing the loss gap, indicating that models fine-tuned with our approach achieve good modality alignment. Secondly, as illustrated in Figure 6 of the main text and Figure 3, our method enables the model to focus on the appropriate scope. In the baseline model, due to its strong visual learning, it tends to focus on the most discriminative small features in the image (second column), which is detrimental to modality alignment. In this case, the visual features are dominated by a small, discriminative part of the image. For example, in the case of "a photo of a Strawberry_healthy," the most discriminative feature of the strawberry leaf is its serrated edges, but this feature alone cannot fully represent the semantic information of "strawberry leaf." Our method, by suppressing this visual learning, leads to a more generalized model. As seen in the fourth column, after applying our method, the model is able to focus correctly on the image features corresponding to the semantic information. This promotes cross-modal alignment, resulting in a significantly higher similarity between the image and text features.

H. Extended Results on CDFSL Task

Table 3 is an extended version of Table 2 in the main text. It includes the performance of various models on four CDFSL datasets under different settings. These settings involve different backbones (CLIP [40], SigLIP2 [46], and PE-Core [3]), the use of a source dataset (Source), and whether fine-tuning on the target domain is applied (FT). Specifically, the methods include ATA [48], AFA [21], wave-SAN [12], StyleAdv [13], StyleAdv-FT (fine-tuned StyleAdv), DARA [65], PMF [20], FLoR [74], CD-CLS [75], AttnTemp [73], VDB [58], IM-DCL [53], StepSTP [54], CoOp [69], Tip-Adapter [63], AMU-Tuning [44], LP++ [22], LDC [30], Maple [25], and CLIP-LoRA [60], which are introduced as our competitors.

I. Few Shot Learning Results

We further evaluate the effectiveness of our method on 11 commonly used classification datasets. Following the setup of previous work [60, 69], we evaluate our approach on 11 datasets covering various fine-grained classification tasks: scenes (SUN397 [51]), aircraft types (Aircraft [36]), satellite imagery (EuroSAT [18]), automobiles (StanfordCars [29]), food items (Food101 [4]), pet breeds (OxfordPets [39]), flowers (Flower102 [38]), general objects (Caltech101 [11]), textures (DTD [7]), and human actions (UCF101 [43]), in addition to ImageNet [9]. These datasets provide a comprehensive benchmarking framework for evaluating few-shot visual classification tasks.

We compare our model against several prompt-based methods, including CoOp [69] with 4 learnable tokens, CoOp [69] with 16 learnable tokens, CoCoOp [68], PLOT++ [6] (an adaptation of the original PLOT designed for transformer architectures), KgCoOp [56], MaPLe [25], and ProGrad [70] with 16 tokens. Additionally, we evaluate adapter-based methods such as Tip-Adapter-F [64] and TaskRes [59] and LDC [30], as well as the LoRA-based method CLIP-LoRA [60]. These comparative methods provide a comprehensive benchmark to assess the effectiveness of our proposed approach in few-shot visual classification tasks.

Our method is plug-and-play, allowing for quick integration into existing approaches. As shown in Table 4, applying our method to the existing CLIP-LoRA [60] model effectively enhances its few-shot classification performance across 11 datasets. In addition, it achieves the highest average classification accuracy under various shot settings.

J. Datasets

Following previous works [40, 54, 58, 71], we do not utilize source domain datasets and finetune our model directly on the target domain. For the target domain we utilize CropDisease [37], EuroSAT [18], ISIC [8], and ChestX [50], which are cross-domain datasets from the domain of agriculture, remote sensing, and medical data with significant domain gaps.

K. Implementation Details

In the experiments, we adopt the ViT-Base/16 network as the primary feature extraction network, with parameters pre-trained by CLIP [40], SigLIP2 [46] and PE-Core [3]. For the parameter λ, we consider two cases: when $\mathcal{L}_{\text{ad}}$ is used for the visual branch (e.g., Maple, CLIP-LoRA), λ is set to 0.1; when used for the text branch (e.g., CoOp), λ is set to 0.001. The parameter β is set to 0.5 in all cases. For each episode, following [54], we first perform data augmentation on the support samples. Subsequently, for each method, we train for 250 epochs, and the first 150 epochs are used as the initial epochs, employing $\mathcal{L}_{\text{ad}}$ and $\mathcal{L}_{\text{ra}}$. We evaluate every network using 15 query samples per class, randomly selecting 800 tasks, and report the average results (%) with the 95% confidence interval. We use an NVIDIA GeForce RTX 3090 GPU for training and testing.

L. Broader Impact

In this paper, we observe the misalignment of modalities in CLIP under cross-domain scenarios and reveal that the cross-entropy-based fine-tuning process inherently incorporates strong visual learning. This learning direction acts as a shortcut, allowing the model to reduce the loss without considering cross-modal relationships, leading to suppression of cross-modal relationship learning. Based on these findings, We then propose two methods to suppress visual learning and enhance cross-modal alignment. Our research is crucial for future studies on fine-tuning VLM models in cross-domain scenarios. It highlights the impact of visual learning on fine-tuning, a factor that has been overlooked in previous work. While our method has been evaluated across four distinct target domains, offering a promising initial assessment of its cross-domain applicability, the diversity of these domains may not fully capture all potential real-world scenarios. Future work will focus on expanding our evaluations to include a broader range of target domains to better understand the method's performance in diverse real-world contexts.

Table 3. The accuracy(%) of four target domain datasets under 5-way 1-shot and 5-way 5-shot tasks. The use of a source dataset (Source), and whether fine-tuning on the target domain is applied (FT)

Task	Method	backbone	Source	FT	ISIC	EuroSAT	CropDisease	ChestX	Avg
5-way 1-shot	ATA [48]	RN10	Y	-	33.21±0.40	61.35±0.50	67.47±0.50	22.10±0.20	46.03
	AFA [21]	RN10	Y	-	33.21±0.30	63.12±0.50	67.61±0.50	22.92±0.20	46.72
	wave-SAN [12]	RN10	Y	-	33.35±0.71	69.64±1.09	70.80±1.06	22.93±0.49	49.18
	StyleAdv [13]	RN10	Y	-	33.96±0.57	70.94±0.82	74.13±0.78	22.64±0.35	50.42
	ATA-FT [48]	RN10	Y	Y	34.94±0.40	68.62±0.50	75.41±0.50	22.15±0.20	50.28
	DARA [65]	RN10	Y	Y	36.42±0.64	67.42±0.80	80.74±0.76	22.92±0.40	51.88
	StyleAdv-FT [13]	RN10	Y	Y	35.76±0.52	72.92±0.75	80.69±0.28	22.64±0.35	53.00
	PMF [20]	ViT/DINO	Y	Y	30.36±0.36	70.74±0.63	80.79±0.62	21.73±0.30	50.91
	StyleAdv-FT [13]	ViT/DINO	Y	Y	33.99±0.46	74.93±0.58	84.11±0.57	22.92±0.32	53.99
	FLoR [74]	ViT/DINO	Y	Y	35.49	73.09	83.55	23.26	53.85
	CD-CLS [75]	ViT/DINO	Y	Y	35.56	74.97	84.53	23.39	54.62
	AttnTemp [73]	ViT/DINO	Y	Y	38.05	75.09	84.78	23.63	55.39
	FN+VDB [58]	RN18	-	Y	32.96±0.57	69.67±0.80	79.68±0.74	22.64±0.41	51.24
	IM-DCL [53]	RN10	-	Y	38.13±0.57	77.14±0.71	84.37±0.99	23.98±0.79	55.91
	StepSTP [54]	ViT/CLIP	-	Y	32.97±0.27	70.01±0.21	84.84±0.72	**22.84±0.95**	52.68
	CoOp [69]	ViT/CLIP	-	Y	32.86±0.47	72.08±0.66	80.50±0.74	21.65±0.32	51.77
	Tip-Adapter [63]	ViT/CLIP	-	Y	32.68±0.37	75.44±0.51	77.15±0.66	22.24±0.26	51.87
	PromptSRC [26]	ViT/CLIP	-	Y	31.86±0.57	73.44±0.71	76.15±0.89	21.16±0.36	50.65
	PDA [1]	ViT/CLIP	-	Y	31.45±0.44	69.68±0.71	79.20±0.83	20.66±0.28	50.25
	AMU-Tuning [44]	ViT/CLIP	-	Y	32.29±0.67	72.24±0.71	80.20±0.86	21.56±0.36	51.57
	LP++ [22]	ViT/CLIP	-	Y	33.63±0.41	73.05±0.55	81.84±0.66	21.72±0.42	52.56
	LDC [30]	ViT/CLIP	-	Y	33.72±0.46	74.39±0.52	84.07±0.61	22.32±0.36	53.62
	Maple [25]	ViT/CLIP	-	Y	33.38±0.49	76.05±0.63	81.78±0.72	21.09±0.31	53.07
	Maple + OURS	ViT/CLIP	-	Y	35.11±0.51	76.92±0.65	82.51±0.69	21.64±0.34	54.05
	CLIP-LoRA-Vision [60]	ViT/CLIP	-	Y	36.40±0.42	81.72±0.52	84.62±0.62	21.86±0.32	56.07
	CLIP-LoRA-Vision + OURS	ViT/CLIP	-	Y	**38.12±0.48**	**85.02±0.46**	**87.20±0.51**	22.68±0.41	**58.26**
	SigLIP2-LoRA [46]	ViT/SigLip2	-	Y	33.47	74.16	87.50	21.44	54.14
	SigLIP2-LoRA + OURS	ViT/SigLip2	-	Y	**36.88**	**78.04**	**90.85**	**22.27**	**57.01**
	PE-Core-LoRA [3]	ViT/PE-Core	-	Y	40.89	84.49	91.75	22.02	59.78
	PE-Core-LoRA + OURS	ViT/PE-Core	-	Y	**45.01**	**86.83**	**93.03**	**23.66**	**62.14**
5-way 5-shot	ATA [48]	RN10	Y	-	44.91±0.40	83.75±0.40	90.59±0.30	24.32±0.40	60.89
	AFA [21]	RN10	Y	-	46.01±0.40	85.58±0.40	88.06±0.30	25.02±0.20	61.17
	wave-SAN [12]	RN10	Y	-	44.93±0.67	85.22±0.71	89.70±0.64	25.63±0.49	61.37
	StyleAdv [13]	RN10	Y	-	45.77±0.51	86.58±0.54	93.65±0.39	26.07±0.37	63.02
	ATA-FT [48]	RN10	Y	Y	49.79±0.40	89.64±0.30	95.44±0.20	25.08±0.20	64.99
	DARA [65]	RN10	Y	Y	56.28±0.66	85.84±0.54	95.32±0.34	27.54±0.42	66.25
	StyleAdv-FT [13]	RN10	Y	Y	53.05±0.54	91.64±0.43	96.51±0.28	26.24±0.35	66.86
	PMF [20]	ViT/DINO	Y	Y	50.12	85.98	92.96	27.27	64.08
	StyleAdv-FT [13]	ViT/DINO	Y	Y	51.23±0.51	90.12±0.33	95.99±0.27	26.97±0.33	66.08
	FLoR [74]	ViT/DINO	Y	Y	53.06	90.75	96.47	27.02	66.83
	CD-CLS [75]	ViT/DINO	Y	Y	54.69	91.53	96.27	27.66	67.54
	AttnTemp [73]	ViT/DINO	Y	Y	54.91	90.82	96.66	28.03	67.61
	FN+VDB [58]	RN18	-	Y	47.48±0.59	87.31±0.50	94.63±0.37	25.55±0.43	64.74
	IM-DCL [53]	RN10	-	Y	52.74±0.69	89.47±0.42	95.73±0.38	28.93±0.41	66.72
	StepSTP [54]	ViT/CLIP	-	Y	52.12±0.36	89.40±1.05	96.01±0.88	26.36±0.97	65.97
	CoOp [69]	ViT/CLIP	-	Y	45.78±0.75	85.88±0.49	93.31±0.57	23.35±0.50	62.08
	Tip-Adapter [63]	ViT/CLIP	-	Y	46.96±0.59	87.24±0.33	94.19±0.39	24.07±0.44	63.12
	PromptSRC [26]	ViT/CLIP	-	Y	46.09±0.48	86.54±0.49	89.97±0.41	23.51±0.47	61.52
	PDA [1]	ViT/CLIP	-	Y	45.19±0.62	86.21±0.44	92.67±0.39	21.87±0.33	61.48
	AMU-Tuning [44]	ViT/CLIP	-	Y	44.60±0.62	88.47±0.39	94.26±0.52	23.34±0.41	62.66
	LP++ [22]	ViT/CLIP	-	Y	48.49±0.44	87.48±0.42	94.47±0.38	23.89±0.29	63.58
	LDC [30]	ViT/CLIP	-	Y	49.70±0.33	90.82±0.22	96.71±0.34	25.89±0.21	65.78
	Maple [25]	ViT/CLIP	-	Y	48.35±0.75	89.04±0.52	93.50±0.54	22.96±0.50	63.46
	Maple + OURS	ViT/CLIP	-	Y	50.01±0.78	91.00±0.50	94.48±0.50	23.45±0.49	64.74
	CLIP-LoRA-Vision [60]	ViT/CLIP	-	Y	52.22±0.71	93.31±0.47	95.88±0.42	24.61±0.47	66.50
	CLIP-LoRA-Vision + OURS	ViT/CLIP	-	Y	**56.14±0.46**	**94.14±0.34**	**96.64±0.39**	**26.61±0.43**	**68.38**
	SigLIP2-LoRA [46]	ViT/SigLip2	-	Y	51.79	91.39	96.43	24.24	65.96
	SigLIP2-LoRA + OURS	ViT/SigLip2	-	Y	**55.12**	**92.10**	**97.37**	**26.44**	**67.43**
	PE-Core-LoRA [3]	ViT/PE-Core	-	Y	58.81	94.07	97.25	24.44	68.64
	PE-Core-LoRA + OURS	ViT/PE-Core	-	Y	**61.41**	**94.83**	**98.13**	**26.77**	**70.29**

Table 4. Detailed results for 11 datasets using ViT-B/16 as the visual backbone are presented. Top-1 accuracy, averaged over 3 random seeds, is reported. The highest value is highlighted in bold, and the second-highest performance is underlined.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
	CoOp(4)	68.0	67.3	26.2	50.9	67.1	82.6	90.3	72.7	93.2	50.1	70.7	67.2
	CoOp(16)	65.7	67.0	20.8	56.4	67.5	84.3	90.2	78.3	92.5	50.1	71.2	67.6
	CoCoOp	69.4	68.7	28.1	55.4	67.6	84.9	91.9	73.4	94.1	52.6	70.4	68.8
	TIP-Adapter-F	69.4	67.2	28.8	67.8	67.1	85.8	90.6	83.8	94.0	51.6	73.4	70.9
	CLIP-Adapter	67.9	65.4	25.2	49.3	65.7	86.1	89.0	71.3	92.0	44.2	66.9	65.7
	PLOT++	66.5	66.8	28.6	65.4	68.8	86.2	91.9	80.5	**94.3**	54.6	74.3	70.7
1	KgCoOp	68.9	68.4	26.8	61.9	66.7	86.4	92.1	74.7	94.2	52.7	72.8	69.6
	TaskRes	69.6	68.1	31.3	65.4	68.8	84.6	90.2	81.7	93.6	53.8	71.7	70.8
	MaPLe	69.7	69.3	28.1	29.1	67.6	85.4	91.4	74.9	93.6	50.0	71.1	66.4
	ProGrad	67.0	67.0	28.8	57.0	68.2	84.9	91.4	80.9	93.5	52.8	73.3	69.5
	LDC	69.5	68.0	27.5	78.4	68.2	85.8	91.3	83.6	93.8	**58.2**	73.2	72.5
	CLIP-LoRA	70.4	70.4	30.2	72.3	70.1	84.3	92.3	83.2	93.7	54.3	76.3	72.5
	CLIP-LoRA + Ours	**70.4**	**70.7**	**30.9**	**78.4**	**70.7**	**86.5**	**92.4**	**83.9**	93.9	54.8	**76.4**	**73.5**
	CoOp(4)	68.7	68.0	28.1	66.2	70.5	82.6	89.9	80.9	93.0	53.7	73.5	70.5
	CoOp(16)	67.0	67.0	25.9	65.1	70.4	84.4	89.9	88.0	93.1	54.1	74.1	70.8
	CoCoOp	70.1	69.4	29.3	61.8	68.4	85.9	91.9	77.8	94.4	52.3	73.4	70.4
	TIP-Adapter-F	70.0	68.6	32.8	73.2	70.8	86.0	91.6	90.1	93.9	57.8	76.2	73.7
	CLIP-Adapter	68.2	67.2	27.0	51.2	66.6	86.2	89.7	71.7	93.4	45.4	68.4	66.8
	PLOT++	68.3	68.1	31.1	76.8	73.2	86.3	92.3	89.8	94.7	56.7	76.8	74.0
2	KgCoOp	69.6	69.6	28.0	69.2	68.2	86.6	92.3	79.8	94.5	55.3	74.6	71.6
	TaskRes	70.2	70.5	32.7	70.2	72.1	85.6	90.7	84.4	94.3	55.6	75.2	72.9
	MaPLe	70.0	70.7	29.5	59.4	68.5	86.5	91.8	79.8	**94.9**	50.6	74.0	70.5
	ProGrad	69.1	69.0	31.1	66.3	72.4	84.8	91.5	87.5	93.6	56.0	75.6	72.4
	LDC	69.8	69.6	30.0	81.7	70.75	86.1	91.2	88.7	94.3	**62.2**	75.9	74.5
	CLIP-LoRA	70.8	71.3	33.2	82.7	73.2	83.2	91.3	89.8	94.6	59.9	**80.0**	75.5
	CLIP-LoRA + Ours	**70.8**	**72.3**	**33.4**	**83.1**	**73.5**	**86.6**	**92.5**	**90.4**	94.7	60.8	79.5	**76.2**
	CoOp (4)	69.7	70.6	29.7	65.8	73.4	83.5	92.3	86.6	94.5	58.5	78.1	73.0
	CoOp (16)	68.8	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.0
	CoCoOp	70.6	70.4	30.6	61.7	69.5	86.3	92.7	81.5	94.8	55.7	75.3	71.7
	TIP-Adapter-F	70.7	70.8	35.7	76.8	74.1	86.5	91.9	92.1	94.8	59.8	78.1	75.6
	CLIP-Adapter	68.6	68.0	27.9	51.2	67.5	86.5	90.8	73.1	94.0	46.1	70.6	67.7
	PLOT++	70.4	71.7	35.3	83.2	76.3	86.5	92.6	92.9	95.1	62.4	79.8	76.9
4	KgCoOp	69.9	71.5	32.2	71.8	69.5	86.9	92.6	87.0	95.0	58.7	77.6	73.9
	TaskRes	71.0	72.7	33.4	74.2	76.0	86.0	91.9	85.0	95.0	60.1	76.2	74.7
	MaPLe	70.6	71.4	30.1	69.9	70.1	86.7	93.3	84.9	95.0	59.0	77.1	73.5
	ProGrad	70.2	71.7	34.1	69.6	75.0	85.4	92.1	91.1	94.4	59.7	77.9	74.7
	LDC	71.0	72.9	31.9	86.3	75.1	86.7	91.8	93.9	95.2	**66.4**	79.7	77.3
	CLIP-LoRA	71.4	72.8	37.9	84.9	77.4	82.7	91.0	93.7	95.2	63.8	81.1	77.4
	CLIP-LoRA + Ours	**71.4**	**73.7**	**39.2**	**86.4**	**78.0**	**87.0**	**93.4**	**94.3**	**95.8**	65.2	**81.3**	**78.7**
	CoOp(4)	70.8	72.4	37.0	74.7	76.8	83.3	92.1	95.0	94.7	63.7	79.8	76.4
	CoOp(16)	70.6	71.9	38.5	77.1	79.0	82.7	91.3	94.9	94.5	64.8	80.0	76.8
	CoCoOp	70.8	71.5	32.4	69.1	70.4	87.0	93.3	86.3	94.9	60.1	75.9	73.8
	TIP-Adapter-F	71.7	73.5	39.5	81.3	78.3	86.9	91.8	94.3	95.2	66.7	82.0	78.3
	CLIP-Adapter	69.1	71.7	30.5	61.6	70.7	86.9	91.9	83.3	94.5	50.5	76.2	71.5
	PLOT++	71.3	73.9	41.4	88.4	81.3	86.6	93.0	95.4	95.5	66.5	82.8	79.6
8	KgCoOp	70.2	72.6	34.8	73.9	72.8	87.0	93.0	91.5	95.1	65.6	80.0	76.0
	TaskRes	72.3	74.6	40.3	77.5	79.6	86.4	92.0	96.0	95.3	66.7	81.6	78.4
	MaPLe	71.3	73.2	33.8	82.8	71.3	87.2	93.1	90.5	95.1	63.0	79.5	76.4
	ProGrad	71.3	73.0	37.7	77.8	78.7	86.1	92.2	95.0	94.8	63.9	80.5	77.4
	LDC	72.4	75.5	38.0	90.8	79.7	86.9	92.5	96.0	95.9	**71.5**	80.9	80.0
	CLIP-LoRA	72.3	74.7	45.7	89.7	82.1	83.1	91.7	96.3	95.6	67.5	84.1	80.3
	CLIP-LoRA + Ours	**72.4**	**75.5**	**48.6**	**90.9**	**82.5**	**87.3**	**94.0**	**96.9**	**95.9**	69.5	**84.1**	**81.6**
	CoOp(4)	71.5	74.6	40.1	83.5	79.1	85.1	92.4	96.4	95.5	69.2	81.9	79.0
	CoOp(16)	71.9	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.0
	CoCoOp	71.1	72.6	33.3	73.6	72.3	87.4	93.4	89.1	95.1	63.7	77.2	75.4
	TIP-Adapter-F	73.4	76.0	44.6	85.9	82.3	86.8	92.6	96.2	95.7	70.8	83.9	80.7
	CLIP-Adapter	69.8	74.2	34.2	71.4	74.0	87.1	92.3	92.9	94.9	59.4	80.2	75.5
	PLOT++	72.6	76.0	46.7	92.0	84.6	87.1	93.6	97.6	96.0	71.4	85.3	82.1
16	KgCoOp	70.4	73.3	36.5	76.2	74.8	87.2	93.2	93.4	95.2	68.7	81.7	77.3
	TaskRes	73.0	76.1	44.9	82.7	83.5	86.9	92.4	97.5	95.8	71.5	84.0	80.8
	MaPLe	71.9	74.5	36.8	87.5	74.3	87.4	93.2	94.2	95.4	68.4	81.4	78.6
	ProGrad	72.1	75.1	43.0	83.6	82.9	85.8	92.8	96.6	95.9	68.8	82.7	79.9
	LDC	**73.8**	76.9	47.8	92.1	84.1	87.3	93.3	97.8	96.4	**77.0**	84.4	82.8
	CLIP-LoRA	73.6	76.1	54.7	92.1	86.3	84.2	92.4	98.0	96.4	72.0	**86.7**	83.0
	CLIP-LoRA + Ours	73.4	**76.9**	**57.3**	**92.5**	**86.6**	**87.7**	**94.5**	**98.4**	**96.5**	73.2	86.1	**84.0**

M. Pseudocode

Below is the pseudocode for fine-tuning the model in a few-shot learning setting. The parts highlighted in red correspond to the implementation of our method. Our method can be implemented with just a few lines of code.

Algorithm 1 Suppressing Visual Learning and Relationship Alignment

Require: Pretrained VLMs with visual encoder θ_v and text encoder θ_t, tokenized prompts for all classes $\{r_k\}$, few-shot labeled samples $\{(\mathbf{x}_i, y_i)\}$, and hyperparameters λ and β.

for each labeled sample $(\mathbf{x}_i, y_i)$ **do**
  Obtain visual feature $f_i = \theta_v(\mathbf{x}_i)$
end for
for each prompt $r_j \in \{r_k\}$ **do**
  Obtain text feature $t_j = \theta_t(r_j)$
end for
Compute logits $\mathcal{S} = \mathcal{F}\mathcal{T}^T$.
Using $\mathcal{S}$ to compute cross-entropy loss $\mathcal{L}_{\text{vlm}}$.
Compute visual feature similarity matrix $A^v = \mathcal{F}\mathcal{F}^T$.
Compute text feature similarity matrix $A^t = \mathcal{T}\mathcal{T}^T$.
Randomly generate index I_{rand}.
Compute class-shffle logits as $\mathcal{S}_{rand} = A^v[:, I_{rand}]$.
Using $\mathcal{S}_{rand}$ as new logits to compute cross-entropy loss as the anti-visual loss $\mathcal{L}_{\text{ad}}$.
Fusing A^v and A^t to get A^{fuse}.
Compute relationship alignment loss $\mathcal{L}_{\text{ra}}$.
Total loss $\mathcal{L} = \mathcal{L}_{\text{vlm}} + \beta\mathcal{L}_{\text{ra}} + \lambda\mathcal{L}_{\text{ad}}$.

References

[1] Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 729–737, 2024. 7

[2] Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27004–27014, 2024. 1

[3] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 3, 6, 7

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 6

[5] Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023. 2

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 1, 6

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[8] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 6

[9] 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. 6

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[12] Yuqian Fu, Yu Xie, Yanwei Fu, Jingjing Chen, and Yu-Gang Jiang. Wave-san: Wavelet based style augmentation network for cross-domain few-shot learning. *arXiv preprint arXiv:2203.07656*, 2022. 1, 6, 7

[13] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24575–24584, 2023. 6, 7

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 2

[15] 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 *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*, pages 124–141. Springer, 2020. 1

[16] Xiaoshuai Hao and Wanqian Zhang. Uncertainty-aware alignment network for cross-domain video-text retrieval. *Advances in Neural Information Processing Systems*, 36:38284–38296, 2023. 2

[17] Masashi Hatano, Ryo Hachiuma, Ryo Fujii, and Hideo Saito. Multimodal cross-domain few-shot learning for egocentric action recognition. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019. 6

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[20] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 6, 7

[21] Yanxu Hu and Andy J Ma. Adversarial feature augmentation for cross-domain few-shot classification. In *European conference on computer vision*, pages 20–37. Springer, 2022. 1, 6, 7

[22] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23773–23782, 2024. 2, 6, 7

[23] Zhong Ji, Zhishen Hou, Xiyao Liu, Yanwei Pang, and Jungong Han. Information symmetry matters: a modal-alternating propagation network for few-shot learning. *IEEE Transactions on Image Processing*, 31:1520–1531, 2022. 2

[24] Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10678–10686, 2020. 2

[25] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1, 6, 7

[26] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200, 2023. 7

[27] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 1

[28] Sanghyeon Kim, Hyunmo Yang, Yunghyun Kim, Youngjoon Hong, and Eunbyung Park. Hydra: Multi-head low-rank adaptation for parameter efficient fine-tuning. *Neural Networks*, page 106414, 2024. 2

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[30] Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25411–25421, 2025. 2, 6, 7

[31] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 1

[32] Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3618–3627, 2025. 1

[33] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9424–9434, 2021. 1

[34] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 2

[35] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. *arXiv preprint arXiv:2311.17091*, 2023. 1

[36] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[37] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:215232, 2016. 6

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[39] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[40] 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. 1, 3, 6

[41] Hossein Rajabzadeh, Mojtaba Valipour, Tianshu Zhu, Marzieh Tahaei, Hyock Ju Kwon, Ali Ghodsi, Boxing Chen, and Mehdi Rezagholizadeh. Qdylora: Quantized dynamic low-rank adaptation for efficient large language model tuning. *arXiv preprint arXiv:2402.10462*, 2024. 2

[42] Zeyu Shangguan, Daniel Seita, and Mohammad Rostami. Cross-domain multi-modal few-shot object detection via rich text. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6570–6580. IEEE, 2025. 2

[43] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[44] Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Amu-tuning: Effective logit bias for clip-based few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23323–23333, 2024. 2, 6, 7

[45] Hongduan Tian, Feng Liu, Zhanke Zhou, Tongliang Liu, Chengqi Zhang, and Bo Han. Mind the gap between prototypes and images in cross-domain finetuning. *Advances in Neural Information Processing Systems*, 37:11251–11289, 2025. 2

[46] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 6, 7

[47] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*, 2022. 2

[48] Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. *arXiv preprint arXiv:2104.14385*, 2021. 1, 6, 7

[49] Haohan Wang, Zeyi Huang, Hanlin Zhang, Yong Jae Lee, and Eric P Xing. Toward learning human-aligned cross-domain robust models by countering misaligned features. In *Uncertainty in Artificial Intelligence*, pages 2075–2084. PMLR, 2022. 2

[50] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 6

[51] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[52] Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[53] Huali Xu, Li Liu, Shuaifeng Zhi, Shaojing Fu, Zhuo Su, Ming-Ming Cheng, and Yongxiang Liu. Enhancing information maximization with distance-aware contrastive learning for source-free cross-domain few-shot learning. *IEEE Transactions on Image Processing*, 2024. 6, 7

[54] Huali Xu, Yongxiang Liu, Li Liu, Shuaifeng Zhi, Shuzhou Sun, Tianpeng Liu, and MingMing Cheng. Step-wise distribution alignment guided style prompt tuning for source-free cross-domain few-shot learning. *arXiv preprint arXiv:2411.10070*, 2024. 1, 6, 7

[55] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23826–23837, 2024. 2

[56] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 1, 6

[57] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 1

[58] Moslem Yazdanpanah and Parham Moradi. Visual domain bridge: A source-free domain adaptation for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2868–2877, 2022. 1, 6, 7

[59] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 6

[60] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 2, 6, 7

[61] Baoquan Zhang, Chuyao Luo, Demin Yu, Huiwei Lin, Xutao Li, Yunming Ye, and Bowen Zhang. Metadiff: Meta-learning with conditional diffusion for few-shot learning. In *AAAI*, 2024. 1

[62] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023. 2

[63] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 6, 7

[64] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 2, 6

[65] Yifan Zhao, Tong Zhang, Jia Li, and Yonghong Tian. Dual adaptive representation alignment for cross-domain few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11720–11732, 2023. 6, 7

[66] Yumiao Zhao, Bo Jiang, Yuhe Ding, Xiao Wang, Jin Tang, and Bin Luo. Fine-grained vlm fine-tuning via latent hierarchical adapter learning. *arXiv preprint arXiv:2508.11176*, 2025. 2

[67] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20061–20070, 2023. 1

[68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1, 6

[69] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 6, 7

[70] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 1, 6

[71] Linhai Zhuo, Zheng Wang, Yuqian Fu, and Tianwen Qian. Prompt as free lunch: Enhancing diversity in source-free cross-domain few-shot learning through semantic-guided prompting. *arXiv preprint arXiv:2412.00767*, 2024. 1, 6

[72] Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*, 2023. 2

[73] Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 1, 6, 7

[74] Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, and Ruixuan Li. Flatten long-range loss landscapes for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23575–23584, 2024. 6, 7

[75] Yixiong Zou, Shuai Yi, Yuhua Li, and Ruixuan Li. A closer look at the cls token for cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:85523–85545, 2025. 1, 6, 7