Supplementary Document for "Language Does Matter for Cross-Domain Few-Shot Visual Feature Enhancement"

The supplementary material presents a series of extended analyses and experimental validations to further support our contributions. Specifically, it includes: 1) a more detailed discussion of related work; 2) in-depth analyses of the motivation for introducing language prompts; 3) additional justification for the simplicity and effectiveness of our design; 4) an analysis contrasting our method with conventional multimodal fusion strategies; and 5) expanded qualitative results, including more visualizations of feature enhancement, semantic segmentation effects, and object detection outcomes.

1. More Detailed Related Work

In-domain FSL. Human agents can quickly perceive and adapt to new situations with only a few examples, a capability rooted in the brain's ability to generalize from accumulated experiences across diverse contexts. Inspired by this form of intelligence, researchers [1–3] have introduced few-shot learning (FSL), a machine learning paradigm designed to overcome the limitations of conventional approaches that rely heavily on large labeled datasets. Existing FSL approaches include metric-based methods[2, 4–8], which construct task-invariant feature spaces for efficient similarity comparisons; optimization-based methods [3, 9–14], which use meta-learning to facilitate rapid adaptation; and transfer learning-based methods [15–21], which fine-tune pre-trained models for downstream tasks.

Cross-domain FSL. In real-world scenarios, few-shot learners often face domain shifts [22–31], where source-domain data consist of well-annotated natural images [2, 32], while deployment domains, such as remote sensing [33] or medical imaging [34, 35], suffer from data scarcity and high annotation costs. To address this, recent studies focus on task-specific fine-tuning using limited support set. Guo et al. [28] fine-tune task-specific classifiers while updating a subset of backbone parameters, whereas Teng et al. [36] refine feature affine transformation layers using few-shot samples. Li et al. [30] propose task-adaptive adapters, learned from support set, to generate task-specific weights for multiple backbone blocks, while Guo et al. [28] leverage reinforcement learning to dynamically adjust feature transformation parameters for each task. Beyond these task-specific adaptations, domain adaptation and regularization strategies have been explored to mitigate domain shifts and enhance generalization. Fu et al. [23] employ adversarial training to eliminate domain-style biases in feature representations, while Zou et al. [37] introduce a novel normalization layer that flattens the loss landscape, facilitating more effective fine-tuning. Perera et al. [29] propose a parameter-efficient linear transformation of pre-trained features to alleviate overfitting during fine-tuning. Additionally, Zou et al. [37] identify that the class token in Vision Transformers [38] (ViTs) encodes source-domain-specific information, which can hinder adaptation; to address this, they propose randomly initializing the class token to mitigate source-domain interference. Recently, prompt-based fine-tuning has emerged as an alternative approach for efficient task adaptation. Zhuo et al. [39] introduce learnable visual prompts to distill semantic knowledge from CLIP [40] while fine-tuning task-specific classification heads. Wu et al. [41] encode task-relevant information from support set into learnable visual prompts, integrating them into the ViT backbone for task-adaptive feature modulation. Wu et al. [42] further extend this approach by storing domain-specific knowledge in learnable prompts, retrieving domain-relevant prompts based on the support set, and fine-tuning them for task-specific adaptation.

While advanced fine-tuning methods improve cross-domain adaptability in FSL, they often rely on limited visual cues for adaptation. These cues primarily capture superficial style or appearance variations in support set, failing to explicitly guide the model toward domain-relevant knowledge. To address this limitation, we present an embarrassingly simple cross-modal visual feature enhancement framework that introduces linguistic descriptions of image attributes to regulate the pre-trained visual feature model for specific target image adaptation. Although prior studies have explored the use of textual modality knowledge to enhance few-shot learning, they primarily focus on constructing semantically enriched prototype representations through multimodal integration. For instance, Han et al. [43] treated class labels in few-shot tasks as textual descriptions, encoding class-level text prototypes with CLIP and fusing them with visual prototypes. Similarly, Shangguan et al. [44] employed text-visual

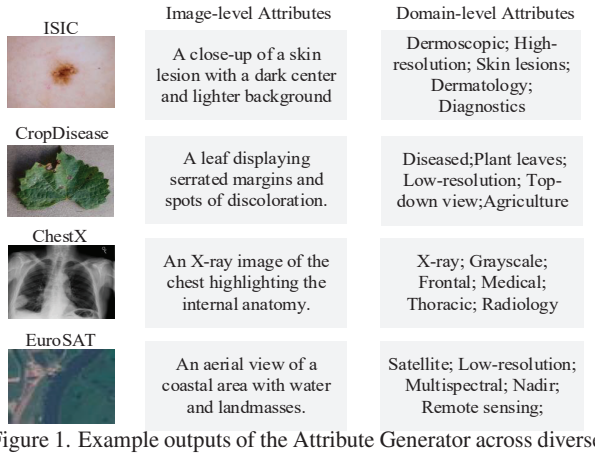


Figure 1. Example outputs of the Attribute Generator across diverse target domains: ISIC (a medical dataset for skin lesion segmentation with high intra-class variation), CropDisease (a collection of crop disease images from various plant species, with challenging variations in environmental conditions), EuroSAT (a remote sensing dataset for land cover classification with fine-grained spatial details), and ChestX (a medical dataset containing chest X-ray images with a variety of pathologies and image quality variations).

fusion to build robust class prototypes, emphasizing the alignment and aggregation of multimodal features through textual knowledge. In contrast, this study focuses on embedding linguistic descriptions of image attributes into the visual feature representation, thus compensating for the limitations of pure visual cues in capturing cross-domain transferable high-level semantic characteristics.

2. Additional Analysis

2.1. Motivation for Introducing Language Prompts

Cross-domain few-shot learning (CD-FSL) poses two core challenges that severely limit the transferability of visual representations. First, the presence of substantial distribution shifts between source and target domains—compounded by large intra-class visual variability and the scarcity of labeled target data—often causes models to overfit to superficial, domain-specific cues. These non-transferable shortcut patterns, while effective within the source domain, fail to generalize to unseen domains, leading to degraded performance. Second, the limited supervision in the target domain restricts the model's ability to disentangle domain-invariant semantics from domain-specific noise, making it difficult to acquire robust and generalizable representations for downstream adaptation.

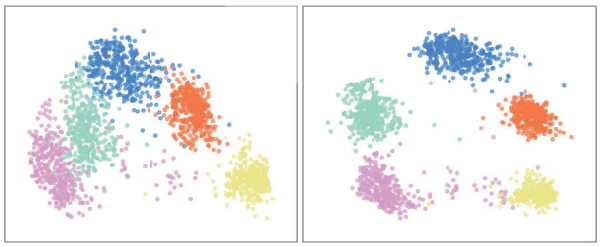
To address these issues, we introduce language prompts as cross-modal semantic priors that serve as complementary guidance to purely visual features. Unlike visual cues, which may implicitly encode brittle or spurious correlations, linguistic descriptions are explicit, semantically structured, and naturally domain-agnostic. Specifically, we extract both fine-grained image-level attributes (e.g., object appearance, color, texture) and coarse-grained domain-level attributes (e.g., scene style, background context) using a pre-trained image captioning model and a large language model. Examples of such domain-level prompts across various challenging target domains are illustrated in Fig. 1. These descriptions are then embedded into the visual feature space via a residual cross-attention mechanism during the adaptation phase, providing structured semantic signals that guide the learning of transferable features. This strategy offers several key benefits: 1) It regularizes the adaptation process by injecting external, high-level semantics that are less sensitive to visual distribution shifts; 2) It helps the model focus on meaningful, invariant patterns rather than task-irrelevant details; 3) It enhances alignment between modality-invariant concepts and domain-specific variations, facilitating more robust cross-domain generalization.

The language prompt mechanism is task-agnostic, plug-and-play, and compatible with off-the-shelf pre-trained visual backbones, making it both practical and broadly applicable across CD-FSL tasks. As shown in Tables 1–3 of the main manuscript, our method achieves consistent and significant gains over strong baselines across a diverse set of benchmarks, including image classification, semantic segmentation, and object detection. These results confirm the effectiveness of using language prompts to inject structured, domain-agnostic semantics, and validate their role in addressing the core limitations of CD-FSL.

2.2. Further Justification for the "Simple yet Effective" Design

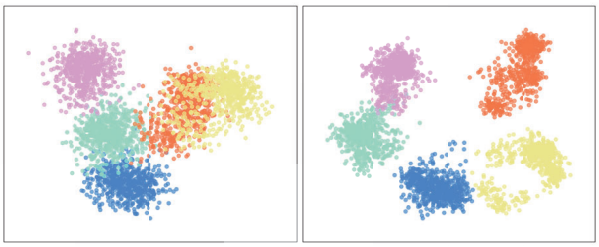
Our residual cross-attention module is purposefully designed as a lightweight and plug-and-play component that can be seamlessly embedded into a wide range of backbone architectures, including Vision Transformers and ResNet. Structurally, it comprises a single cross-attention layer augmented with a residual connection, allowing linguistic descriptions of image attributes to be embedded directly into the visual feature space without disrupting the original architecture. This minimalist design preserves the integrity of the pre-trained visual backbone while enabling effective semantic alignment between modalities. Importantly, the module introduces only a negligible number of additional parameters and requires no modification to the backbone, making it exceptionally easy to integrate and highly adaptable.

In contrast to existing approaches that often rely on complex multi-branch fusion structures, heavy prompt tuning pipelines, or tightly coupled architecture-specific designs, our method offers a generalizable and implementation-friendly solution. Its simplicity not only reduces design and training complexity but also enhances portability across tasks and model families—an especially valuable property for cross-domain few-shot scenarios where resource constraints and domain shifts pose significant challenges.



(a) StyleAdv (b) StyleAdv+Ours

Figure 2. Visualization of feature distribution on the EuroSAT dataset.



(a) StyleAdv (b) StyleAdv+Ours

Figure 3. Visualization of feature distribution on the ISIC dataset.

We further substantiate the effectiveness of this simple design through extensive empirical studies. As shown in Tables 1–3 of the main manuscript, our framework consistently improves performance across a variety of tasks, including cross-domain few-shot classification, semantic segmentation, and object detection. These gains hold across multiple challenging benchmarks and strong baselines, highlighting the broad applicability and strong generalization capability of our method. Moreover, efficiency analysis (Table 6) shows that the residual cross-attention module incurs only marginal computational overhead, confirming that its simplicity does not come at the cost of practicality. Together, these results validate that a well-designed, minimally invasive mechanism for embedding structured linguistic information can yield substantial improvements in performance—underscoring the central claim of our framework as "simple yet effective".

2.3. Regarding Query-specific Linguistic Representations at Inference Time

During inference, the query-specific linguistic representations are generated solely from the individual query image by leveraging a pre-trained image captioning model and a large language model to produce semantic attribute descriptions. This process is analogous to a forward pass on a single sample, similar to standard feature extraction. The resulting semantic attributes are then used to refine the query's visual features via a residual cross-attention mechanism, enriching the semantic content of the query representation.

Importantly, our inference procedure does not involve simultaneous use of multiple query samples, nor does it leverage information from other queries or the overall query distribution. This clearly distinguishes our method from transductive inference approaches in CD-FSL, which adapt the model or features by exploiting the entire query set in an unsupervised manner. In contrast, our method functions strictly in an inductive, per-sample fashion, preserving the independence of each query and upholding the integrity of the few-shot evaluation protocol. Accordingly, the inclusion of query-specific semantic information constitutes a per-query feature refinement that is fully consistent with the standard assumptions of CD-FSL.

In conclusion, although our method leverages semantic attributes at query time, it fully respects the foundational principles and constraints of the CD-FSL, achieving improved cross-domain generalization without compromising the fairness or validity of evaluation.

3. Additional Visualization

3.1. More Visualization of Feature Enhancement

We present additional visual analyses in Fig.2 and Fig.3. These t-SNE plots vividly illustrate the contrasting distribution patterns of features extracted by our method versus the baseline.

Concretely, the baseline features demonstrate significant overlap between different classes, indicating weak inter-class separability and ambiguous decision boundaries. This feature entanglement often stems from the model's reliance on superficial or domain-specific cues that do not generalize well when faced with domain shifts. In stark contrast, features produced by our method form compact, well-separated clusters, evidencing enhanced intra-class cohesion and clearer inter-class distinctions.

This marked improvement is primarily driven by the integration of linguistically enriched attribute descriptions through our residual cross-attention mechanism. By embedding structured, high-level semantic information into the visual feature space, the model is encouraged to attend to semantically meaningful and robust cues, rather than noise or spurious visual artifacts. As a result, the learned representations better capture intrinsic category semantics and demonstrate greater resilience to domain variation.

In addition, this semantic augmentation aids in disentangling domain-invariant features from domain-specific noise—a critical advantage in few-shot cross-domain learning scenarios where labeled target data is limited. Collectively, these visualizations offer compelling qualitative evidence that our approach not only sharpens feature discriminability but also facilitates the acquisition of transferable representations, thereby substantially boosting generalization performance across diverse downstream tasks and domains.

3.2. Visualization of Segmentation Effects

To qualitatively demonstrate the performance gains of our proposed method over the baseline IFA [45] in semantic segmentation tasks, we visualize segmentation results on several representative few-shot tasks. Specifically, we randomly select a 1-shot task from each CD-FSS dataset, fine-tune the model using the corresponding support image, and evaluate on the associated query images. The qualitative results are presented across four benchmark datasets: ISIC, Chest X-Ray, FSS-1000, and Deepglobe, shown respectively in Fig.4, Fig.5, Fig.6, and Fig.7.

From these visualizations, several clear advantages of our method over the baseline emerge. First, our approach more accurately captures semantic regions, enabling more complete and coherent foreground segmentation, as exemplified in Fig.4 and Fig.6. This suggests improved semantic understanding and feature representation robustness. Second, our method produces noticeably sharper and more precise segmentation boundaries, particularly evident in Fig.5, reflecting enhanced spatial localization ability and boundary awareness. Third, it demonstrates stronger robustness in modeling complex semantic relationships, effectively reducing misclassifications and segmentation errors in challenging scenarios such as those shown in Fig.7.

Collectively, these qualitative results offer strong evidence that our proposed method significantly enhances semantic segmentation performance across a variety of datasets, demonstrating its capability to produce more accurate, robust, and generalizable segmentation outcomes.

3.3. Visualization of Detection Effects

To qualitatively demonstrate the detection performance of our proposed method, we visualize representative detection results on multiple diverse datasets, including ArTaxOr (Fig.8), Clipart1k (Fig.9), DeepFish (Fig.10), and UODD (Fig.11). For each dataset, several query images were randomly selected, and the detection outputs of both our method and the CD-ViTO baseline are presented alongside the corresponding ground truth annotations. Different colors are used to distinguish the detected bounding boxes from each method, enabling clear visual comparison of their respective detection capabilities.

Our method consistently produces more precise bounding boxes that better conform to object shapes, effectively reducing under- and over-detection errors. This advantage is particularly evident in challenging scenes involving cluttered backgrounds, small or partially occluded objects, and diverse object appearances, such as those found in the DeepFish and UODD datasets. Moreover, our approach demonstrates superior recall, successfully identifying difficult and ambiguous instances that the baseline often misses or detects incompletely, as seen in the Clipart1k and ArTaxOr datasets.

Beyond improvements in localization and recall, our method exhibits stronger robustness in distinguishing true objects from background clutter and visually similar distractors. This capability leads to fewer false positives and more confident detections, producing cleaner and more reliable results closely aligned with ground truth. The reduction in spurious detections highlights the effectiveness of the cross-modal semantic guidance embedded in our framework.

Importantly, these visual results underscore the strong cross-domain generalization ability of our method across domains with significant visual differences, ranging from natural and medical images to cartoons, underwater scenes, and aerial photography. By incorporating rich linguistic semantic cues into visual representations, our framework effectively bridges domain gaps and mitigates distribution shifts, enabling robust few-shot detection performance in diverse and challenging target domains.

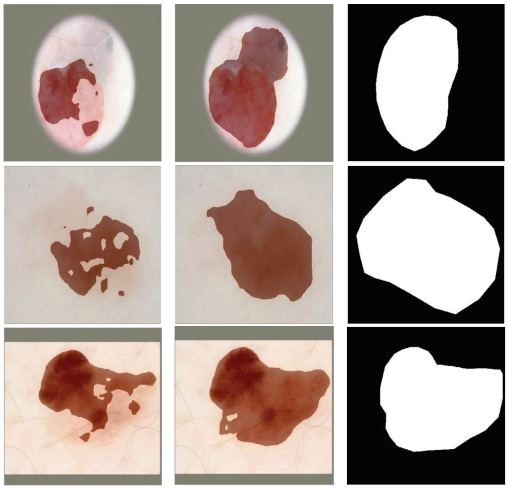
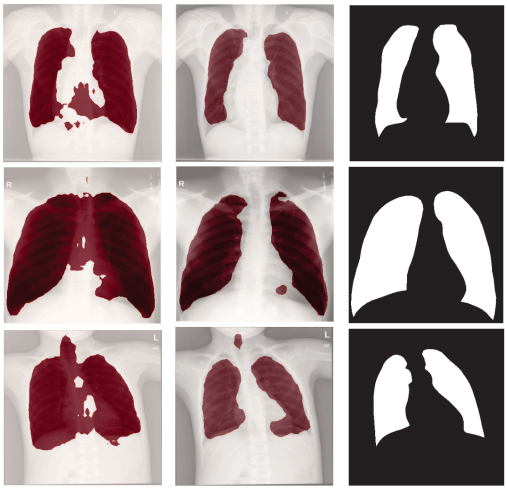
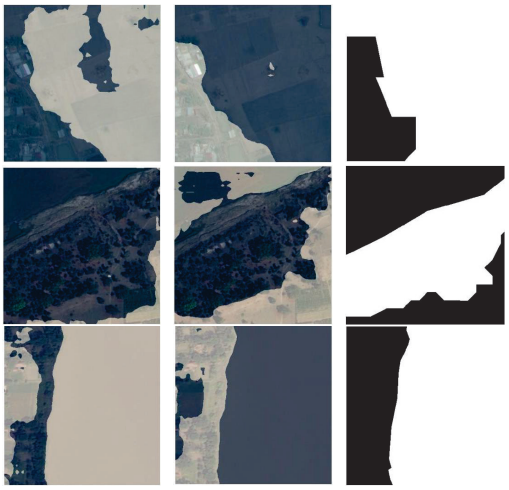


Figure 4. Visualization of segmentation effects on ISIC dataset.



IFA IFA+Ours Ground Truth

Figure 5. Visualization of segmentation effects on Chest X-Ray dataset.

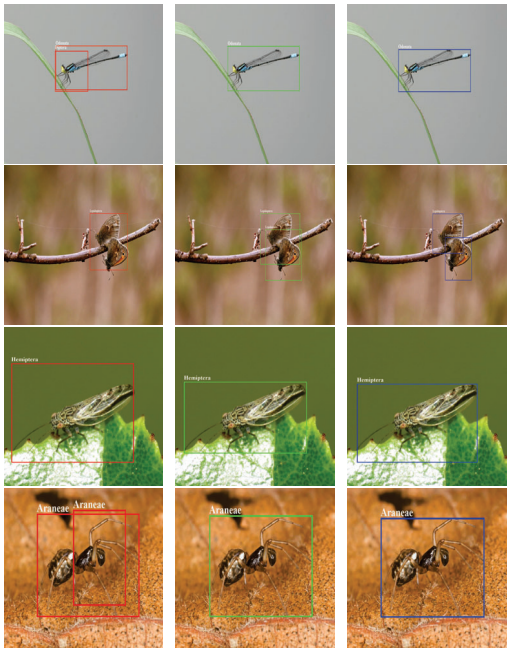


IFA IFA+Ours Ground Truth

Figure 7. Visualization of segmentation effects on Deepglobe dataset.

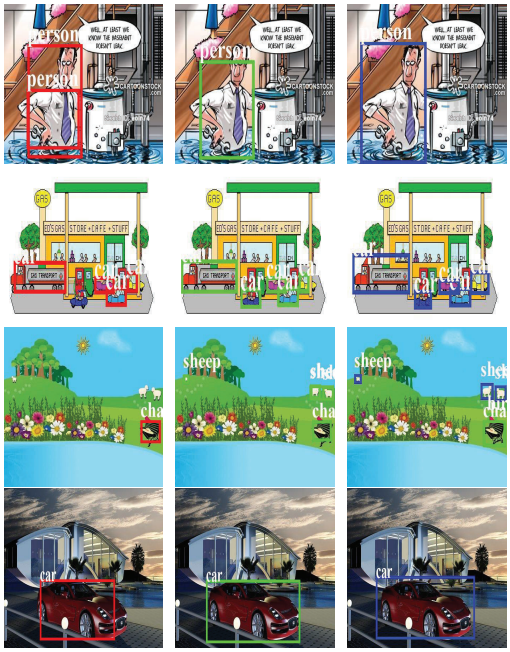
IFA IFA+Ours Ground Truth

Figure 6. Visualization of segmentation effects on FSS-1000 dataset.



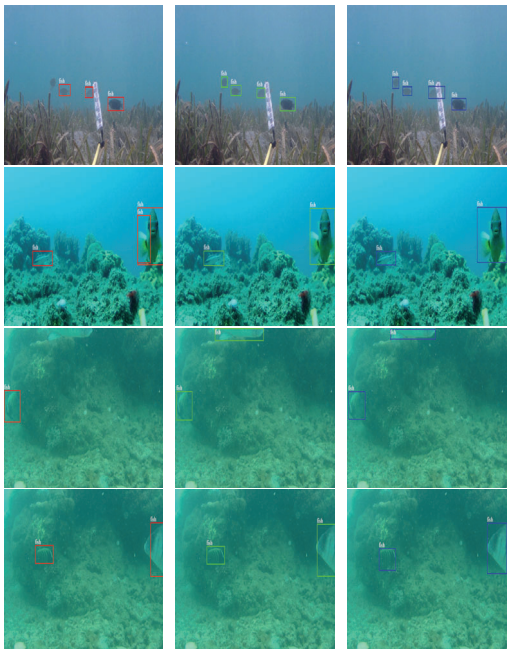
CD-ViTO CD-ViTO+Ours Ground Truth

Figure 8. Visualization of detection effects on ArTaxOr dataset.



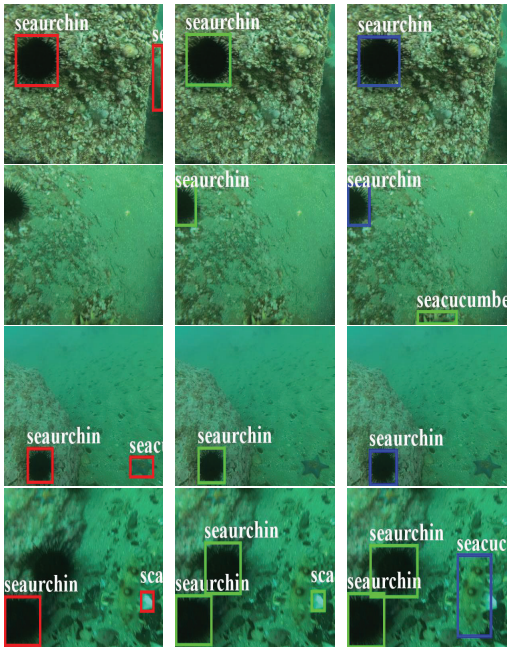
CD-ViTO CD-ViTO+Ours Ground Truth

Figure 9. Visualization of detection effects on Clipart1k dataset.



CD-ViTO CD-ViTO+Ours Ground Truth

Figure 10. Visualization of detection effects on DeepFish dataset.



CD-ViTO CD-ViTO+Ours Ground Truth

Figure 11. Visualization of detection effects on UODD dataset.

References

[1] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 1

[2] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 1

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 1

[4] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille, 2015. 1

[5] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[6] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[7] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020.

[8] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7972–7981, 2022. 1

[9] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. 1

[10] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

[11] Baoquan Zhang, Chuyao Luo, Demin Yu, Xutao Li, Huiwei Lin, Yunming Ye, and Bowen Zhang. Metadiff: Meta-learning with conditional diffusion for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16687–16695, 2024.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[13] Siyuan Sun and Hongyang Gao. Meta-adam: An meta-learned adaptive optimizer with momentum for few-shot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Nishant Jain, Arun S Suggala, and Pradeep Shenoy. Improving generalization via meta-learning on hard samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27600–27609, 2024. 1

[15] 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

[16] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023.

[17] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023.

[18] Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36:55361–55374, 2023.

[19] 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.

[20] Shuai Shao, Yu Bai, Yan Wang, Baodi Liu, and Bin Liu. Collaborative consortium of foundation models for open-world few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4740–4747, 2024.

[21] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23681–23690, 2024. 1

[22] 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

[23] 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. 1

[24] Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1075–1081. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[25] Yi Wu, Ziqiang Li, Chaoyue Wang, Heliang Zheng, Shanshan Zhao, Bin Li, and Dacheng Tao. Domain re-modulation for few-shot generative domain adaptation. volume 36, 2024.

[26] 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.

[27] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22584–22591, 2024.

[28] Yurong Guo, Ruoyi Du, Yuan Dong, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Task-aware adaptive learning for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1590–1599, 2023. 1

[29] Rashindrie Perera and Saman Halgamuge. Discriminative sample-guided and parameter-efficient feature space adaptation for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23794–23804, 2024. 1

[30] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7161–7170, 2022. 1

[31] Hao Zheng, Runqi Wang, Jianzhuang Liu, and Asako Kanezaki. Cross-level distillation and feature denoising for cross-domain few-shot classification. *arXiv preprint arXiv:2311.02392*, 2023. 1

[32] Alex Krizhevsky, I Sutskever, and G Hinton. Imagenet classification with deep convolutional neural. In *Neural Information Processing Systems*, pages 1–9, 2014. 1

[33] 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. 1

[34] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1

[35] 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. 1

[36] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2019. 1

[37] 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

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1

[39] 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

[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

[41] Jiamin Wu, Xin Liu, Xiaotian Yin, Tianzhu Zhang, and Yongdong Zhang. Task-adaptive prompted transformer for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6012–6020, 2024. 1

[42] Jiamin Wu, Tianzhu Zhang, and Yongdong Zhang. Hybridprompt: Domain-aware prompting for cross-domain few-shot learning. *International Journal of Computer Vision*, 132(12):5681–5697, 2024. 1

[43] Guangxing Han, Long Chen, Jiawei Ma, Shiyuan Huang, Rama Chellappa, and Shih-Fu Chang. Multi-

modal few-shot object detection with meta-learning-based cross-modal prompting. *arXiv preprint arXiv:2204.07841*, 2022. 1

[44] Zeyu Shangguan, Daniel Seita, and Mohammad Rostami. Cross-domain multi-modal few-shot object detection via rich text. *arXiv preprint arXiv:2403.16188*, 2024. 1

[45] Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. Pixel matching network for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 978–987, 2024. 4