Supplementary Document for "Towards Effective Foundation Model Adaptation for Extreme Cross-Domain Few-Shot Learning"

This supplementary document includes:

- Providing a more detailed discussion of related works in Sec. 1.
- Including additional implementation details in Sec. 2.
- Revisiting the adaptation of foundation models for extreme cross-domain few-shot tasks in Sec. 3.
- Clarifying the motivation for utilizing the expert model in Sec. 4.1.
- Explaining why not use the expert model directly to solve the target task in Sec. 4.2
- Clarifying how our approach avoids the potential overfitting risk of the expert model in Sec. 4.3
- Verifying whether our approach improves the expert model in Sec. 4.4.
- Validating smaller expert models in Sec. 4.5.
- Providing more visualizations in Sec. 4.6.
- Evaluating hyper-parameters in Sec. 4.7.

1. More Related Work

Few-shot learning. Few-shot learning [1] (FSL) aims to generalize the model to target tasks using only a limited number of samples. Previous works in few-shot learning can be categorized into two main pipelines: meta-learning [2–8] and transfer learning [9–14]. Among them, meta-learning methods focus on extracting shared knowledge between different tasks to assist the model in handling novel tasks. To achieve this goal, meta-learning methods explicitly organize the training data into the form of support-query tasks. During this process, the support set is utilized for constructing inner optimization, while the query set is used to evaluate the generalization error of the current model and perform outer optimization. By sequentially solving thousands of training tasks, the model acquires the ability to capture shared knowledge across tasks. Transfer learning methods typically adopt a two-stage paradigm, comprising pre-training and fine-tuning. During the pre-training phase, the model is fed with a large amount of base classes data to help it acquire initial representation capabilities. Subsequently, the model undergoes task-level fine-tuning to adapt to specific target tasks. Despite making some progress, these few-shot learning methods are difficult to apply in cross-domain scenarios [15, 16].

Cross-domain Few-shot learning. To address the challenges posed by domain shift, recent works [16–25] have developed few-shot learners capable of cross-domain transfer. Zhou et al. [17] posit that the simplicity bias of global features constitutes a primary constraint on model cross-domain generalization. To address this, they propose a global-local semantic alignment framework to aid the model in learning comprehensive representations. Fu et al. [18] introduce an adversarial training framework to mitigate the interference of image styles on model generalization. Similarly, Wang et al.[19] and Hu et al.[21] propose constructing adversarial training at the task-level and feature-level, respectively, to assist the model in learning robust features. Zheng et al. [25] introduce a cross-level knowledge distillation approach to extract features effectively. Zou et al.[26] examine the loss landscape in representation spaces, identifying sharp minima as critical obstacles to both knowledge transfer and fine-tuning, and introduce a normalization layer to smooth these minima over longer distances. Wu et al.[27] combine domain-specific and adaptive prompts to support domain-aware knowledge transfer, improving generalization in unfamiliar domains. Zhao et al.[28] tackle cross-domain few-shot learning (CD-FSL) by using a dual adaptive representation alignment method, enhancing fast model adaptation by aligning prototype features and normalized distributions. Wang et al.[29] propose a meta-memory framework to address domain shifts, integrating style memory for richer feature distributions and content memory for semantic information, effectively bridging domain gaps and improving cross-domain performance. Perera et al.[30] offer a lightweight, parameter-efficient adaptation approach alongside a discriminative, sample-aware loss function to improve cross-domain generation. Yang et al.[31] approach domain discrepancies and noisy data through progressive training combined with adaptive distillation, yielding more robust performance in the face of challenging domain shifts.

Although these methods have achieved state-of-the-art performance, they are generally constrained by the use of shallow backbone networks (e.g., ResNet-10) to learn feature initialization models from small-scale source domains like mini-ImageNet. In the age of large models, this labor-intensive source domain training process becomes redun-

dant. With the availability of open-source foundation models that offer robust feature initialization, the emphasis of cross-domain few-shot learning should now shift towards the efficient adaptation of foundation models to target domains. As a result, this study concentrates on addressing the challenges of adapting foundation models in extreme cross-domain scenarios.

Foundation models and adaptation strategies. Thanks to the remarkable scalability of Transformers [32] and the advancement of self-supervised learning techniques, several outstanding foundation models are gradually emerging. A notable example is DINO [33], which learns transferable visual feature representations through contrastive learning on large-scale unlabeled datasets. Additionally, Masked AutoEncoders (MAE) [34] randomly mask patches within images and then predict the masked regions from the unmasked ones, enabling the model to learn semantic features of the images. The CLIP[35] maps images and text into the same representation space and is trained by contrasting the similarity and dissimilarity between different pairs of images and text, thereby learning feature representations with strong generalization capabilities.

On the other hand, adapting foundation models to downstream tasks is a newly emerging research direction. Some research [10, 36–38, 38–42, 42, 43] seek to effectively leverage foundation models in downstream few-shot tasks, aim to fully harness the benefits of pre-trained knowledge while minimizing the data needed for effective fine-tuning. Typically, LoRA [44] achieves model fine-tuning solely by adding additional low-rank matrices. Visual Prompt Tuning (VPT) [45] achieves model adaptation by introducing a small number of trainable parameters into the input space. Zhou et al. [36] propose to automate CLIP prompt engineering by utilizing learnable context vectors, aiming to improve the efficiency of few-shot transfer learning. Song et al. [40] propose to use lightweight residual adapters to fine-tune CLIP features, enabling efficient online few-shot learning without requiring offline fine-tuning. CLIP-Adapter[41] improves vision-language models by employing feature adapters and residual blending during fine-tuning, outperforming traditional methods. CO3[42] introduces a combination of frozen foundation models and tunable adapters to integrate pre-trained knowledge, addressing label noise and enhancing few-shot learning performance. Lin et al.[37] investigate the role of cross-modal knowledge to enhance adaptability in few-shot tasks. Zhang et al.[39] introduce a cascade approach that combines CLIP [35] and DINO[33] models to capture a wider range of knowledge, enhancing both image feature extraction and task-specific learning. Silva et al.[38] propose to merge class-adaptive linear probing with enhanced Lagrangian optimization, facilitating efficient knowledge transfer from large pre-trained models.

Although these methods have been successful, they are predominantly tailored for within-domain few-shot tasks. When applied to cross-domain few-shot learning, they face considerable difficulties due to large domain shifts [15, 16]. In contrast, this study centers on adapting foundation models to extreme cross-domain scenarios, proposing a novel adaptation framework designed to improve the generalization ability of pre-trained vision foundation models. This framework enhances the model's performance by incorporating task-relevant knowledge from an expert model during fine-tuning, enabling the foundation model to better manage significant domain shifts in few-shot learning tasks.

2. More Implementation Details

For extreme setting, our experimental dataset includes ISIC [46], Chest [47], EuroSAT [48], and CropDisease [49]. The ISIC dataset consists of medical dermoscopic images, encompassing seven different types of skin diseases. The Chest dataset comprises X-ray images, totaling seven different categories. The EuroSAT dataset consists of remote sensing images collected by the European Space Agency, encompassing ten types of remote sensing scenes. The CropDisease dataset consists of agricultural plant images, encompassing thirty-eight categories, commonly used for plant disease recognition. In this study, we organize these datasets into more challenging All-way K-shot tasks. The number of training and testing examples is presented in Table 1. In terms of optimization, the batch size is set to 32, the learning rate to 0.0001, and training is carried out for 500 epochs. In each target domain, a subset not overlapping with the test samples is designated for model selection. For the hyper-parameters λ_1 and λ_2, we perform a grid search from the candidate set $0.000001, 0.05, 1.0, 10.0, 50.0, 100.0$. The masking ratio is fixed at 0.5. To ensure a fair comparison, we apply the same experimental settings across all baselines and comparison methods.

Table 1. Data partitioning under extreme CD-FSL settings.

	ISIC	Chest	EuroSAT	CropDisease
Classes	7	7	10	38
Training samples for each class	1 or 5	1 or 5	1 or 5	1 or 5
Total test samples	6132	25157	16200	32612

For traditional CD-FSL setting, we follow the benchmarks proposed by Guo et al. [15] and Tseng et al. [50] for experimentation. We rigorously adhere to the data settings of traditional cross-domain few-shot learning to ensure fairness. All comparative results are extracted from the respective papers. Besides, we clarify that our method does not necessitate model pre-training on the source domain dataset. Instead, it emphasizes task-level training for each few-shot task directly within the target domain. Specifically, we train the model using the support set for each sample task and

evaluate its performance on the query set. For each target domain, following the implementation of [15, 17, 18, 51], we conduct experiments by randomly sampling 600 N-way K-shot 15-query few-shot tasks, and measure the average accuracy across these tasks as the evaluation metric. For each few-shot task, we set the number of training epochs to 100 and the batch size to 4. The hyper-parameters λ_1 and λ_2 are both set to 1. Since there is no validation set for model selection, we use the model from the last training epoch for evaluation.

For in-domain few-shot learning, the previous methods first training the model on the base classes of the in-domain dataset (e.g., mini-ImageNet [52], tiered-ImageNet [53], CIFAR-FS [54].). Subsequently, the model is tested on new classes using a few-shot sampling strategy, which generally includes randomly sampling 600 N-way K-shot tasks. For each few-shot task, the support set is utilized for model fine-tuning and classifier construction, after which testing is conducted on the query set. The average performance across all few-shot tasks is taken as the final performance metric for the dataset. To ensure fairness, we maintain the same data settings as previous works [55–57]. For training, we also adhere to the configurations established in our traditional cross-domain few-shot learning setting.

All experiments are conducted on an NVIDIA 3090 GPU using the PyTorch framework.

3. Revisiting Foundation Models Adaptation for Extreme Cross-Domain Few-shot Tasks

Details. We revisit some advanced adaptation strategies to explore the performance of foundation models in extreme cross-domain few-shot learning. We select three representative foundation models, namely DINO [33], Supervised ViT (S-ViT)[32], and Masked AutoEncoders (MAE)[34]. We revisit adaptation methods in four aspects: feature inference, adding a classification head, parameter-efficient fine-tuning, and full-parameter standard fine-tuning. Among these baselines, Prototype inference [2] involves directly utilizing the foundation model for feature extraction, then constructing prototypes using the k-shot samples from each class, and finally using these prototypes to predict the class affiliation for test samples. Linear probing refers to learning a task-specific linear classifier on top of the features extracted from the foundation model. Additionally, we adopt the low-rank adaptation (LoRA) method [44] for parameter-efficient fine-tuning. Standard fine-tuning refers to utilizing few-shot samples from the target domain to perform full-parameter updating of the foundation model.

Results. We validate three representative foundation models (DINO, S-ViT, MAE) under 1-shot and 5-shot settings on four datasets. The experimental results are presented in Table 2, Table 3, Table 4, and Table 5. We can draw the following observations from the experimental results. Firstly, simple feature reuse baselines such as Prototype inference and Linear probing struggle to perform effectively in this extreme cross-domain scenario. This contrasts sharply with the performance of the foundation models under in-domain settings. This suggests that when distributional differences are significant, simple feature reuse methods may fail to harness the potential of the foundation model. Secondly, LoRA fine-tuning performs even worse than linear probing. This suggests that adjusting only a subset of parameters is an inappropriate strategy, as it struggles to correct foundation model's excessive memorization of the original distribution. Additionally, the configuration of such adapters involves very few learnable parameters, which may exacerbate the risk of over-fitting on few-shot samples. Thirdly, standard fine-tuning achieves the best average performance on both 1-shot and 5-shot settings across all four datasets. This indicates that standard fine-tuning remains the most powerful baseline method. Nevertheless, the performance of standard fine-tuning remains sub-optimal.

The reason may lie in the hug domain gap and limited annotated examples in the target task prevents properly adapting the foundation model to fit the task-specific semantic characteristics. To alleviate this dilemma, one possible solution is to exploit the pure task-specific knowledge (e.g., model trained from scratch using the limited annotated samples in the target task) to guide the fine-tuning process of the foundation model. Following this idea, we present a novel absorption adaptation learning framework, which attempt to enhance the extreme cross-domain generalization capacity of a pre-trained vision foundation model (e.g., "F-ViT"). This framework achieves this by through absorbing some task-relevant knowledge from an expert model (e.g., "T-ViT") during fine-tuning. As a result, our method emerges as a front-runner, achieving significant performance improvements. For instance, on the ISIC dataset, our method outperforms the strongest baseline across three foundation models by an average gain of 6.81% (1-shot) and 5.28% (5-shot).

4. Further Analysis

4.1. Why utilize expert model?

The motivation for incorporating an expert model into our framework arises from both theoretical and empirical insights. 1) Theoretically, the expert model, trained exclusively on the target data, acquires pure task-specific knowledge, untainted by unrelated domain features present in foundation model pre-training. This domain-specific focus enables the expert model to emphasize critical patterns relevant to the target task, effectively guiding the foundation model to refine its feature extraction toward task-specific modes. Thus, the expert model functions as a focused advisor, complement-

Table 2. Results on ISIC dataset. The best results are in bold.

	DINO		S-ViT		MAE		Ave.	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Prototype inference	24.14	34.06	14.72	28.86	11.19	30.27	16.68	31.06
Linear probing	24.94	51.88	21.23	33.66	12.19	51.98	19.45	45.83
LoRA fine-tuning	24.44	50.29	18.45	36.21	20.38	47.95	21.09	44.81
Stand fine-tuning	26.35	60.95	20.89	39.82	26.82	57.72	24.68	52.83
Ours	**37.49**	**62.90**	**24.30**	**44.19**	**32.68**	**67.25**	**31.49**	**58.11**

Table 3. Results on Chest dataset. The best results are in bold.

	DINO		S-ViT		MAE		Ave.	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Prototype inference	16.43	19.11	18.34	16.79	19.23	13.50	18.00	16.46
Linear probing	18.29	25.29	17.03	17.57	18.08	28.89	17.80	23.91
LoRA fine-tuning	18.29	25.35	15.85	22.09	17.99	24.89	17.37	24.11
Stand fine-tuning	20.05	26.41	17.23	18.65	21.27	27.18	19.51	24.08
Ours	**21.36**	**28.74**	**20.70**	**28.40**	**22.90**	**29.44**	**21.65**	**28.86**

ing the broader, generalized scope of the foundation model, which, while comprehensive, may overlook domain-specific nuances critical to task performance. 2) Empirically, we observe that the expert and foundation models yield complementary feature representations (as shown in Fig. 1). While the foundation model provides a robust initialization with minimal noise, it sometimes fails to capture key task-specific semantic regions. The expert model, on the other hand, persistently identifies relevant patterns unique to the target task, albeit with some noise. This interplay suggests that by carefully integrating both models, we can harness the specificity of the expert model alongside the stability and generalizability of the foundation model, achieving a balanced and enhanced representation suited for CD-FSL tasks. 3) However, integrating the models effectively presents a significant challenge. Conventional approaches such as knowledge distillation or feature fusion proved insufficient (as presented in Table 6 and Table 7), with the foundation model's generalization tendency often overpowering the expert model's task-specific contributions. To address this, we introduce two novel strategies: masked cross-model unidirectional reconstruction, which selectively injects task-relevant knowledge while tempering the foundation model's propensity for overgeneralization, and decision graph association, which aligns similarity matrices across models to ensure consistency at the decision level, facilitating coherent knowledge transfer. Extensive experiments confirm that our approach improves CD-FSL performance, underscoring the expert model's role in enhancing task-specific insights and validating the effectiveness of our integration techniques.

4.2. Why not use the expert model directly to solve the target task?

Relying solely on the expert model for the target task presents significant challenges due to its inherent limitations. Although the expert model effectively captures task-specific information, it often learns features that are susceptible to noise and lacks the robustness required for reliable generalization. Nevertheless, it remains a valuable component, as it can provide critical insights—such as low-level semantic features and higher-order structural cues—that enhance the understanding of the target task. To leverage these strengths while addressing its limitations, we propose two novel approaches: masked cross-model unidirectional reconstruction and decision graph association. These methods not only mitigate noise in the expert model's output but also enhance its integration with the foundation model, resulting in a balanced and complementary framework that improves overall task performance.

4.3. How to avoid the potential over-fitting risk of expert model?

Our method mitigates the potential over-fitting risk of the expert model by employing high-order semantic structure alignment, i.e., the decision graph association. This approach enforces a structured alignment between the expert and foundation models, focusing not only on individual features but also on their interrelationships within a broader semantic context. By ensuring that the knowledge transfer from the expert model to the foundation model is both balanced and refined, our method reduces the risk of over-fitting that might arise if the models were aligned based solely on isolated features.

In contrast to traditional knowledge distillation methods,

Table 4. Results on EuroSAT dataset. The best results are in bold.

	DINO		S-ViT		MAE		Ave.	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Prototype inference	59.05	78.70	37.29	61.01	34.48	36.23	43.60	58.64
Linear probing	**60.38**	83.02	45.41	61.12	33.72	40.15	46.50	61.43
LoRA fine-tuning	60.08	83.09	44.96	62.61	33.56	38.95	46.19	61.54
Stand fine-tuning	60.24	82.98	47.34	65.93	41.49	64.05	49.69	70.98
Ours	60.28	**83.94**	**48.42**	**69.06**	**43.25**	**65.34**	**50.65**	**72.78**

Table 5. Results on CropDisease dataset. The best results are in bold.

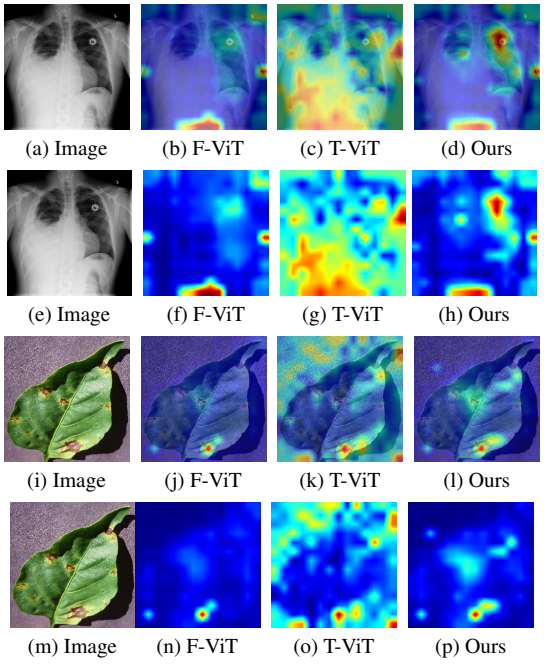
	DINO		S-ViT		MAE		Ave.	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Prototype inference	39.39	73.40	37.62	66.71	7.91	15.14	28.30	51.75
Linear probing	51.01	84.07	31.92	56.71	13.60	17.48	32.17	52.75
LoRA fine-tuning	52.81	84.07	28.30	56.85	16.06	24.15	32.39	55.02
Stand fine-tuning	52.18	86.48	37.73	**66.98**	21.08	57.74	36.99	70.40
Ours	**54.27**	**86.70**	**41.49**	66.53	**21.26**	**58.17**	**39.00**	**70.46**

which transfer the expert model's knowledge to the foundation model in a more direct manner—often leading to over-fitting due to the noisy, task-specific features of the expert model—our approach introduces a more holistic and structured alignment. This enables effective integration of task-specific knowledge without overwhelming the foundation model with redundant or irrelevant information.

Experimental results confirm that, unlike conventional distillation approaches, our method significantly reduces the over-fitting risk by promoting a more cohesive integration of task-specific knowledge, thereby enhancing generalization and performance on the target task.

4.4. Whether our approach improves the expert model?

In our framework, the foundation model and the expert model collaborate through joint training, bridged by two core approaches: masked cross-model unidirectional reconstruction (MCR) and decision graph association (DGA). MCR explicitly injects the task-specific knowledge of the expert model into the foundation model, while DGA achieves cross-model mutual regularization by aligning the higher-order semantic structures between models. With the assistance of the DGA module, the foundation model is also able to transfer knowledge to the expert model, thereby enhancing the quality of the expert model. To validate this, we conducted experiments on the ISIC dataset under the *All*-way K-shot setting. In addition to evaluating 1-shot and 5-shot scenarios, we also explore cases with more shots, as the expert model's quality is expected to improve with increased samples. These experiments aim to validate that our method is effective not only in few-shot scenarios but also in traditional large-sample settings.



(a) Image (b) F-ViT (c) T-ViT (d) Ours
(e) Image (f) F-ViT (g) T-ViT (h) Ours
(i) Image (j) F-ViT (k) T-ViT (l) Ours
(m) Image (n) F-ViT (o) T-ViT (p) Ours

Figure 1. Complementary observations between F-ViT and T-ViT.

The experimental results are shown in Table 8. We can draw two conclusions. First, our approach not only improves the performance of the foundation model but also enhances the expert model. Second, as the number of shots increases, the quality of T-ViT improves. Meanwhile, the complementarity between T-ViT and F-ViT persists, enabling our method to continue yielding benefits.

Table 6. Comparing ours with distillation based methods.

	ISIC		Chest		EuroSAT		CropDisease		Ave.	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Feature-level Distillation	27.76	62.31	**22.01**	27.26	56.28	79.81	53.58	86.12	39.90	63.87
Logits-level Distillation	27.53	61.54	20.67	25.99	**60.88**	83.18	**54.87**	86.51	40.98	64.30
Ours	**37.49**	**62.90**	21.36	**28.74**	60.28	**83.94**	54.27	**86.70**	**43.35**	**65.57**

Table 7. Comparing ours with fusion based methods.

	ISIC		Chest		EuroSAT		CropDisease		Ave.	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Cross-attention Fusion	29.34	62.19	20.78	26.73	58.74	82.40	53.71	85.64	40.64	64.24
Decision Fusion	25.44	56.96	19.85	25.50	58.94	82.76	36.29	61.98	35.13	56.80
Ours	**37.49**	**62.90**	**21.36**	**28.74**	**60.28**	**83.94**	**54.27**	**86.70**	**43.35**	**65.57**

Table 8. Evaluating the improvement of our method over the expert model baseline.

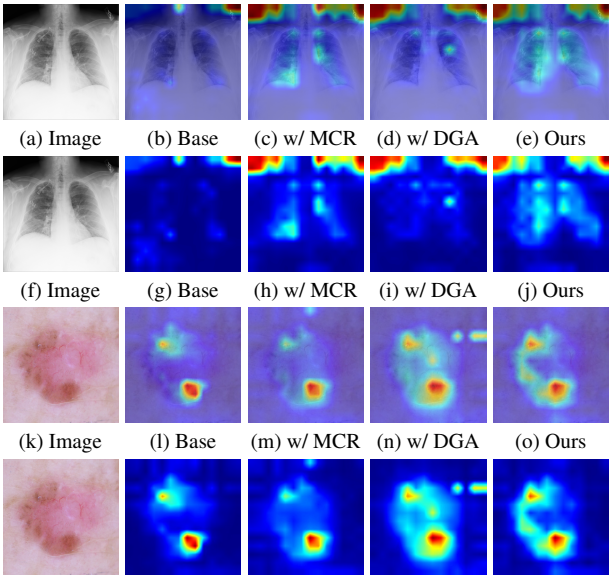
Method	**1-shot**	**5-shot**	**30-shot**	**300-shot**
T-ViT baseline	26.71	53.68	58.34	75.32
T-ViT+Ours	28.64	56.15	61.12	76.43
F-ViT baseline	26.35	60.95	69.43	85.56
F-ViT+Ours	37.49	62.90	70.56	86.07

4.5. Can smaller expert models be used?

In our framework, the selection of the expert model is flexible, with any Vision Transformer architecture theoretically suitable. To validate this, we fixed the foundation model as ViT-Base and selected smaller ViTs as expert models for the experiments. All experiments were conducted on the ISIC dataset under the All-way 1-shot setting. As shown in Table 9, our method consistently outperforms the baselines across different expert models. However, when the expert model is too small, such as ViT-Tiny, the performance of our method significantly declines. In contrast, using ViT-Base as the expert model yields optimal performance, albeit with higher computational costs. In conclusion, the choice of expert model is flexible, with performance gains showing a linear relationship to the model's capacity. Higher performance gains typically come with increased computational overhead, meaning the expert model should be chosen based on available resources in practical deployments.

Table 9. Evaluating smaller T-ViT while keeping F-ViT as ViT-B.

Exepert Model	**ViT-Tiny**	**ViT-S**	**ViT-B**
T-ViT baseline	21.15	27.42	26.71
F-ViT baseline	26.35	26.35	26.35
Ours	26.56	34.41	37.49

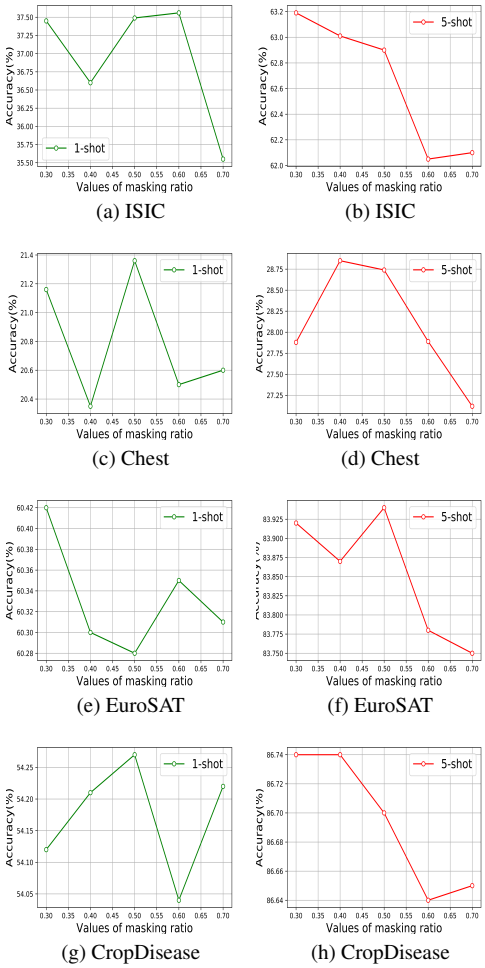


(a) Image (b) Base (c) w/ MCR (d) w/ DGA (e) Ours
(f) Image (g) Base (h) w/ MCR (i) w/ DGA (j) Ours
(k) Image (l) Base (m) w/ MCR (n) w/ DGA (o) Ours
(p) Image (q) Base (r) w/ MCR (s) w/ DGA (t) Ours

Figure 2. Feature visualization for baseline and our method.

4.6. Visualization

Complementary observations. We compare feature highlights to visually illustrate the complementary characteristics between pre-trained vision foundation model and task-specific expert model. The results are depicted in Fig. 1. Among them, "F-ViT" represents the original foundation model without being trained on the target domain, while "T-ViT" represents the task-specific expert model trained from scratch using a small number of samples from the target domain. In each figure, the first row represents the highlighted image, while the second row displays the corresponding feature activation map. We can draw the following observations. Firstly, although the foundation model can provide good initial features, it does not comprehensively cover



(a) ISIC (b) ISIC

(c) Chest (d) Chest

(e) EuroSAT (f) EuroSAT

(g) CropDisease (h) CropDisease

Figure 3. Classification accuracy w.r.t values of masking ratio.

all the target attributes. Secondly, the task-specific model can capture potential properties relevant to the target task, but it may also introduce undesirable noise, leading to poor generalization. Comparing "F-ViT" and "T-ViT", we can observe that their feature activations exhibit complementarity. This motivates us to invent a novel complementary collaborative learning framework, which endows the model with cross-domain generalization capabilities by amalgamating the unique strengths of both the foundation model and task-specific expert model. The visualization results align with our expectations, demonstrating that our proposed approach can integrate the knowledge of both "F-ViT" and "T-ViT", resulting in a robust response, such as clearer activation regions and more intense activation values.

Visual contrasts. We employed the Rollout technique [58] for visual analysis. We compare the proposed method with the baseline case, and also demonstrate the effect of using individual components. The visualization results are depicted in Fig. 2. We can derive the following intuitive understanding. Firstly, compared to the baseline, our method can highlight a broader range of feature areas. This suggests that our method accurately captures the underlying patterns of the target, providing a solid foundation for subsequent classifier decisions. Therefore, our method demonstrates better generalization, resonating with the previous quantitative results. Secondly, when our core components (e.g., MCR, DGA) are used independently, our method can also outperform the baseline case. This observation further confirms the contribution of the proposed components. Additionally, there is diversity in feature activation among different components. This indicates that the proposed MCR and DGA approaches can function independently while also integrating effectively, thereby creating a powerful and cohesive framework.

4.7. Hyper-parameter validation

We investigate the impact of the mask ratio hyper-parameter, exploring values within the range $0.3, 0.4, 0.5, 0.6, 0.7$. The experimental results are presented in Fig. 3. The findings indicate that variations in the masking ratio between 0.3 and 0.7 lead to different, albeit minor, changes in performance across the four datasets. Therefore, unless specified otherwise, we adopt a uniform mask ratio of 0.5 for all datasets in our experiments.

References

[1] Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir Gandomi. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Computing Surveys*, 56(12):1–41, 2024. 1

[2] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 1, 3

[3] 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.

[4] Tao Zhang and Wu Huang. Kernel relative-prototype spectral filtering for few-shot learning. In *European Conference on Computer Vision*, pages 541–557. Springer, 2022.

[5] 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.

[6] 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.

[7] Andrey Zhmoginov, Mark Sandler, and Maksym Vladymyrov. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. In *International Conference on Machine Learning*, pages 27075–27098. PMLR, 2022.

[8] 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. 1

[9] Ilia Sucholutsky and Thomas L Griffiths. Alignment with human representations supports robust few-shot learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 73464–73479, 2023. 1

[10] Samyadeep Basu, Shell Hu, Daniela Massiceti, and Soheil Feizi. Strong baselines for parameter-efficient few-shot fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11024–11031, 2024. 2

[11] Zhiqiang Shen, Zechun Liu, Jie Qin, Marios Savvides, and Kwang-Ting Cheng. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9594–9602, 2021.

[12] Meiqi Sun, Zhonghan Zhao, Wenhao Chai, Hanjun Luo, Shidong Cao, Yanting Zhang, Jenq-Neng Hwang, and Gaoang Wang. Uniap: Towards universal animal perception in vision via few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5008–5016, 2024.

[13] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8635–8643, 2021.

[14] Xingyu Zhu, Shuo Wang, Jinda Lu, Yanbin Hao, Haifeng Liu, and Xiangnan He. Boosting few-shot learning via attentive feature regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7793–7801, 2024. 1

[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 *European conference on computer vision*, pages 124–141. Springer, 2020. 1, 2, 3

[16] 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. 1, 2

[17] 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, 3

[18] 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, 3

[19] 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. 1

[20] 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.

[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

[22] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9526–9535, 2021.

[23] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5326–5334, 2021.

[24] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explanation-guided training for cross-domain few-shot classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7609–7616. IEEE, 2021.

[25] Hao ZHENG, Runqi Wang, Jianzhuang Liu, and Asako Kanezaki. Cross-level distillation and feature denoising for cross-domain few-shot classification. In *The Eleventh International Conference on Learning Representations*. 1

[26] 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. 1

[27] Jiamin Wu, Tianzhu Zhang, and Yongdong Zhang. Hybridprompt: Domain-aware prompting for cross-domain few-shot learning. *International Journal of Computer Vision*, pages 1–17, 2024. 1

[28] 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. 1

[29] Wenjian Wang, Lijuan Duan, Yuxi Wang, Junsong Fan, and Zhaoxiang Zhang. Mmt: cross domain few-shot learning via meta-memory transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[30] 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

[31] Yongjin Yang, Taehyeon Kim, and Se-Young Yun. Leveraging normalization layer in adapters with progressive learning and adaptive distillation for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16370–16378, 2024. 1

[32] 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. 2, 3

[33] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[34] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3

[35] 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. 2

[36] 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. 2

[37] 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. 2

[38] 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. 2

[39] 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. 2

[40] 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. 2

[41] 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

[42] 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. 2

[43] 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. 2

[44] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2, 3

[45] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2

[46] 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. 2

[47] 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. 2

[48] 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. 2

[49] Punam Bedi, Pushkar Gole, and Sumit Kumar Agarwal. 18 using deep learning for image-based plant disease detection. *Internet Of things and machine learning in agriculture*, pages 369–402, 2021. 2

[50] 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. 2

[51] Pan Li, Shaogang Gong, Chengjie Wang, and Yanwei Fu. Ranking distance calibration for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9108, 2022. 3

[52] 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. 3

[53] 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. 3

[54] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. 3

[55] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Holger Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*. 3

[56] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. *Advances in Neural Information Processing Systems*, 35:3582–3595, 2022.

[57] Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, and Wenqiang Zhang. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9119–9129, 2022. 3

[58] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020. 7