One Meta-tuned Transformer is What You Need for Few-shot Learning

Xu Yang [1] **Huaxiu Yao** [2] **Ying Wei** [3]

Abstract

Pre-trained vision transformers have revolutionized few-shot image classification, and it has been recently demonstrated that the previous common practice of meta-learning in synergy with these pre-trained transformers still holds significance. In this work, we design a new framework centered exclusively on attention mechanisms, called MetaFormer, which extends the vision transformers beyond patch token interactions to encompass relationships between samples and tasks simultaneously for further advancing their downstream task performance. Leveraging the intrinsical property of ViTs in handling local patch relationships, we propose Masked Sample Attention (MSA) to efficiently embed the sample relationships into the network, where an adaptive mask is attached for enhancing task-specific feature consistency and providing flexibility in switching between few-shot learning setups. To encapsulate task relationships while filtering out background noise, Patch-grained Task Attention (PTA) is designed to maintain a dynamic knowledge pool consolidating diverse patterns from historical tasks. MetaFormer demonstrates coherence and compatibility with off-the-shelf pre-trained vision transformers and shows significant improvements in both inductive and transductive few-shot learning scenarios, outperforming state-of-the-art methods by up to 8.77% and 6.25% on 12 in-domain and 10 cross-domain datasets, respectively.

1. Introduction

There has been a sustained focus on few-shot learning (Vinyals et al., 2016b; Snell et al., 2017), a paradigm where only a few labeled samples (*a.k.a.* support samples) are provided to predict unlabelled samples (*a.k.a.* query samples). The overarching objective is to emulate human-level intelligence capable of swiftly assimilating new concepts. Meta-learning (Thrun & Pratt, 2012) has been a de-facto approach in dealing with few-shot learning, via leveraging the knowledge learned from previous tasks (Finn et al., 2017; Raghu et al., 2019). State-of-the-art meta-learning practices have evolved from modelling the relationships between samples (Vinyals et al., 2016b; Snell et al., 2017; Oreshkin et al., 2018; Hou et al., 2019; Ye et al., 2020; Chen et al., 2021a) to exploring fine-grained relationships among the building blocks of samples (Zhang et al., 2020a; Doersch et al., 2020; Kang et al., 2021; Afrasiyabi et al., 2022) and, more recently, to conditional meta-learning where the transferable knowledge shared among only closely related tasks improves generalization (Rusu et al., 2018; Yao et al., 2019; Requeima et al., 2019; Bateni et al., 2020; Zhou et al., 2021a; Jiang et al., 2022).

The majority of the aforementioned research advancements have predominantly unfolded within the realm of Convolutional Neural Networks (CNNs) (He et al., 2016; Zagoruyko & Komodakis, 2016). However, recent strides in pre-trained Vision Transformers (ViTs) (Caron et al., 2021; Touvron et al., 2021; Zhou et al., 2022a) have notably surpassed CNNs across a diverse array of vision tasks (Dosovitskiy et al., 2020; Liu et al., 2021; Zhu et al., 2020; Ranftl et al., 2021; Strudel et al., 2021). Consequently, one anticipates that the various levels of relationships delineated earlier can collaborate with ViTs to propel the performance of few-shot learning to new heights. While we acknowledge the preliminary efforts (Hiller et al., 2022; Hu et al., 2022), demonstrating that meta-learning can effectively synergize with pre-trained ViTs to further enhance their few-shot learning performance, it is noteworthy that, to the best of our knowledge, there has been no exploration into employing self-attention as the exclusive building block to model the aforementioned three levels of relationships.

In this work, we pose the question of whether it is plausible to construct a high-performing transformer that simultaneously encapsulates the three distinct relationships. We argue that such a design holds the potential to surmount two inherent limitations in prior CNN-based meta-learning models. First, constructing the correspondence between building blocks, such as optimal matching (Zhang et al., 2020a), in

[1]City University of Hong Kong [2]University of North Carolina at Chapel Hill [3]Nanyang Technological University. Correspondence to: Xu Yang < xyang337-c@my.cityu.edu.hk>, Ying Wei <ying.wei@ntu.edu.sg>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

CNNs remains computationally costly. In contrast, ViTs, leveraging patches as tokens, intrinsically and efficiently establish the relationship between patches as building blocks. Second, the relationships between tasks in existing CNN-based models hinge on task embeddings (Rusu et al., 2018; Achille et al., 2019; Yao et al., 2019), often derived from image-level embeddings. This approach introduces irrelevant noise such as background and compromises local structures. Henceforth, tasks with similar embeddings modulate their parameters (e.g., through a FiLM layer (Perez et al., 2018; Requeima et al., 2019; Triantafillou et al., 2021)) to stay close, while the high dimensionality of deep neural networks renders such modulation less effective to push away dissimilar tasks that inherently need to be distinctly separated. Conversely, ViTs present an alternative approach by modeling task relationships explicitly through self-attention.

Motivated by these observations, we propose a few-shot learning architecture centered exclusively on attention mechanisms. Our framework, named MetaFormer (derived from Meta-tuned Transformer), extends the ViT by augmenting the self-attention mechanism beyond patch token interactions to encompass relationships between samples and tasks. In the context of an N-way K-shot task with n patches per sample, and N support and M query samples, the computation of similarities between all samples incurs a potential time complexity of $\mathcal{O}((n(N+M)K)^2)$. This computational complexity is considerable, particularly in light of the often substantial value of M and the large number of patches (n). To address this challenge, we propose Masked Sample Attention (MSA), which implements sample relationships separately after each attention layer of the original ViT in a decoupled manner. Simultaneously, it involves an adaptive mask to not only enforce consistency in task-specific discriminative features across all samples within a task but also facilitate the switch between few-shot learning setups. For addressing task relationships at the building block level, we propose the Patch-grained Task Attention (PTA). PTA learns a probe vector for each task summarizing discriminative patch-level patterns and maintains a dynamic knowledge pool consolidating diverse patterns from historical tasks. Subsequently, PTA conditions on attention-based mechanisms for knowledge retrieval and aggregation.

The contributions of this integrated transformer are outlined as follows. (1) *Coherence and compatibility*: we provide a coherent few-shot learning framework that exclusively leverages attention mechanisms, thus ensuring compatibility with off-the-shelf ViT-backed pre-trained transformers (including CLIP) and enhancing their few-shot learning performances. (2) *Flexibility*: the adaptive mask surprisingly supports both inductive and transductive few-shot learning protocols. (3) *Practical efficacy*: we conduct our experiments on 12 in-domain and 10 cross-domain few-shot generalization datasets, on which MetaFormer outperforms both inductive and transductive state-of-the-art methods with improvements of up to 8.77% and 6.25%.

2. Related work

Meta-Learning in Few-Shot Classification. Meta-learning serves as a fundamental framework for few-shot learning with the aim of distilling and transferring prior knowledge for quickly adapting to new unseen tasks. Most related to our work are metric-based and feedforward-adaptation meta-learning methods. Metric-based methods (Vinyals et al., 2016b; Snell et al., 2017; Sung et al., 2018; Oreshkin et al., 2018; Lee et al., 2019; Chen et al., 2021a; Ma et al., 2021c; Simon et al., 2020; Lai et al., 2022) seek to embed samples into global universal feature representations and aim at improving distance metrics for better measuring sample similarity in the embedding space. Recent works establish fine-grained local feature matching by leveraging Earth Mover's Distance (Zhang et al., 2020a), spatial prototypes (Doersch et al., 2020) or set-to-set metrics (Afrasiyabi et al., 2022). However, fixed embedding is not very robust and sufficient to accommodate tasks with significant shifts due to cluttered backgrounds and intricate scenes. Several feedforward approaches are proposed to perform task adaptation. One line of work generates task-conditioned classifiers (Rusu et al., 2018; Xu et al., 2020), convolution kernels (Ma et al., 2021b; Zhmoginov et al., 2022), or batch normalization parameters (Requeima et al., 2019; Bateni et al., 2020). Another line directly adapts the feature embedding to new tasks utilizing within-support (Ye et al., 2020) and support-query (Hou et al., 2019; Kang et al., 2021) sample relationships. Our approach follows the second vein and the main difference is that we embed intra-task interactions into the network at various scales to fully leverage coarse-to-fine multi-scale information for learning a richer task-specific feature embedding. To handle tasks with different distributions, a handful of works built upon the gradient-based methods try to extract the underlying task structure for customizing initialization (Yao et al., 2019; 2020; Zhou et al., 2021a; Jiang et al., 2022). However, these algorithms rely on time-consuming clustering and the discriminative task representations are difficult to learn (Jiang et al., 2022). Recent works (Wang et al., 2022a; Douillard et al., 2022; Wang et al., 2022b; Smith et al., 2023) propose inter-task attention to prevent catastrophic forgetting in the continual learning setting. Unlike them retaining a single vector (Wang et al., 2022a), our method maintains a knowledge pool to organize structured meta-knowledge (i.e., key feature patterns), which is then tailored for reuse in the current task through attention-based aggregation.

Vision Transformers for Few-Shot Learning. Large-scale pre-trained vision transformers (Dosovitskiy et al., 2020; Liu et al., 2021; Tu et al., 2022), utilizing the data-driven

prior and self-attention mechanism to encode long-range dependency in the data, generalize better than CNNs (Zhang et al., 2022). Combined with self-supervised pre-training techniques, recent works have shown the potential of few-shot vision transformers via designing self-distillation training to mitigate the overfitting issue (He et al., 2022b; Dong et al., 2022; Lin et al., 2023). For example, HcT (He et al., 2022b) utilize the DINO-based (Caron et al., 2021) teacher-student framework to distill the global class token and train three cascaded transformers with two pooling layers in between. To further supervise the patch tokens, SUN (Dong et al., 2022) adopts the patch-level pseudo labels generated by the teacher network and SMKD (Lin et al., 2023) introduces the patch reconstruction loss in Masked Image Modeling (MIM) (He et al., 2022a; Bao et al., 2022; Zhou et al., 2022a). Beyond self-supervised ViTs, the initial investigations conducted by Hu et al. (Hu et al., 2022) and Hiller et al. (Hiller et al., 2022) underscore the advantageous impact of integrating meta-learning techniques to further augment the few-shot learning capabilities of ViTs. FewTURE (Hiller et al., 2022), specifically, uses a support-aware patch importance mask learned in the inner loop to address supervision collapse. However, its utility is mostly confined to the top classifier, restricting the network's ability to refine features for tasks with significant distribution shifts. We also ground our proposed method on pre-trained vision transformers and introduce a novel meta-learning approach that embeds sample and task relationships into ViTs, enabling the learning of more discriminative features for each specific task. Our contribution is orthogonal to SKMD and we empirically show that MetaFormer can further improve the joint performance.

3. MetaFormer for Few-shot Classification

We present our approach in this section. The overall architecture of our MetaFormer is illustrated in Figure 1. We start by briefly introducing the meta-tuning practice and the self-attention of vision transformer in Section 3.1 and then elaborate on our proposed Masked Sample Attention (MSA) and Patch-grained Task Attention (PTA) in Section 3.2 and Section 3.3, respectively. Finally, using MSA and PTA as core building blocks, we present a new vision transformer with holistic attention for meta-learning in Section 3.4.

3.1. Preliminaries

Meta-tuning Practice. Meta-learning aims to learn a model capable of adapting quickly to recognize new classes with only a few labeled examples. Following recent literature (Hu et al., 2022; Hiller et al., 2022), we initially pre-train a model on the base training dataset and then meta-tune the MetaFormer in the episodic training manner (Vinyals et al., 2016b). In a classical N-way K-shot setting, each episode randomly selects N classes to form the support set $\mathcal{S} = \{(x_i^c, y_i^c)\}_{i=1}^{N\times K}$ containing K samples in each class and the query set $\mathcal{Q} = \left\{\left(x_j^t, y_j^t\right)\right\}_{j=1}^{M}$ with M samples.

Self-attention in ViTs. Given a N-way K-shot task with $NK + M$ images $x \in \mathbb{R}^{Height\times Width\times 3}$ as input, ViTs (Dosovitskiy et al., 2020) first divide individual images into n non-overlapping patches and then map them into d-dimension tokens through a linear projection layer. After that, a trainable class token is prepended as the final input token sequence $\mathbf{X} \in \mathbb{R}^{(NK+M)\times L\times d}$ ($L = n+1$), taken by several Multi-head Self-Attention (MHSA) layers and MLP layers for feature extraction. Consider a MHSA layer with H heads, and query, key, and value embeddings of the input $\mathbf{X}$ are given as $\mathbf{Q} = \mathbf{W}^Q\mathbf{X}$, $\mathbf{K} = \mathbf{W}^K\mathbf{X}$, $\mathbf{V} = \mathbf{W}^V\mathbf{X}$, respectively. The output of MHSA is given as:

$$\begin{aligned} &\mathrm{MHSA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \mathrm{Concat}\left(\mathrm{h}_1,\ldots,\mathrm{h}_H\right)\mathbf{W}^O \\ &\text{where } \mathrm{h}_i = \sigma\left(\mathbf{A}_i\right)\mathbf{V}_i = \sigma\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_k}}\right)\mathbf{V}_i \end{aligned} \tag{1}$$

where $\mathbf{W}^O$ is the output projection matrix, $d_k = d/H$ denotes the head dimension, and $\sigma(\cdot)$ signifies the softmax activation function. The matrix $\mathbf{A}_i \in \mathbb{R}^{L\times L}$ serves as the attention matrix, measuring the pairwise token affinities across different patch locations. Following the MHSA as defined in (1), each token representing a building block within the image gains awareness of its relational context with all other tokens.

3.2. Masked Sample Attention

For computational efficiency, the very few existing methods capture sample relationships by attaching one or several extra attention layers on top of the fixed feature extractor (Ye et al., 2020; Doersch et al., 2020; Hiller et al., 2022). However, it is demonstrated that different layers of the backbone yield different semantic levels of feature embedding and thus different types of knowledge (Raghu et al., 2021). Motivated by this, we propose to leverage coarse-to-fine multi-scale information across layers to capture discriminative sample interactions at the patch token level.

A straightforward and intuitive approach is to perform self-attention over both patch and sample dimensions simultaneously. Given the task input $\mathbf{X}$, the core computation of one MHSA layer primarily revolves around calculating the attention matrix $\mathbf{A}_J \in \mathbb{R}^{(NK+M)L\times(NK+M)L}$ in (1). Therefore, the complexity of the joint space-sample attention is $O((NK+M)^2L^2)$. Such joint space-sample interaction empowers the vision transformer to capture sample relationships for task-specific embedding, but it comes at a high computational cost and incurs heavy memory footprints.

Inspired by the divided space-time attention in video transformers (Bertasius et al., 2021), we introduce a more ef-

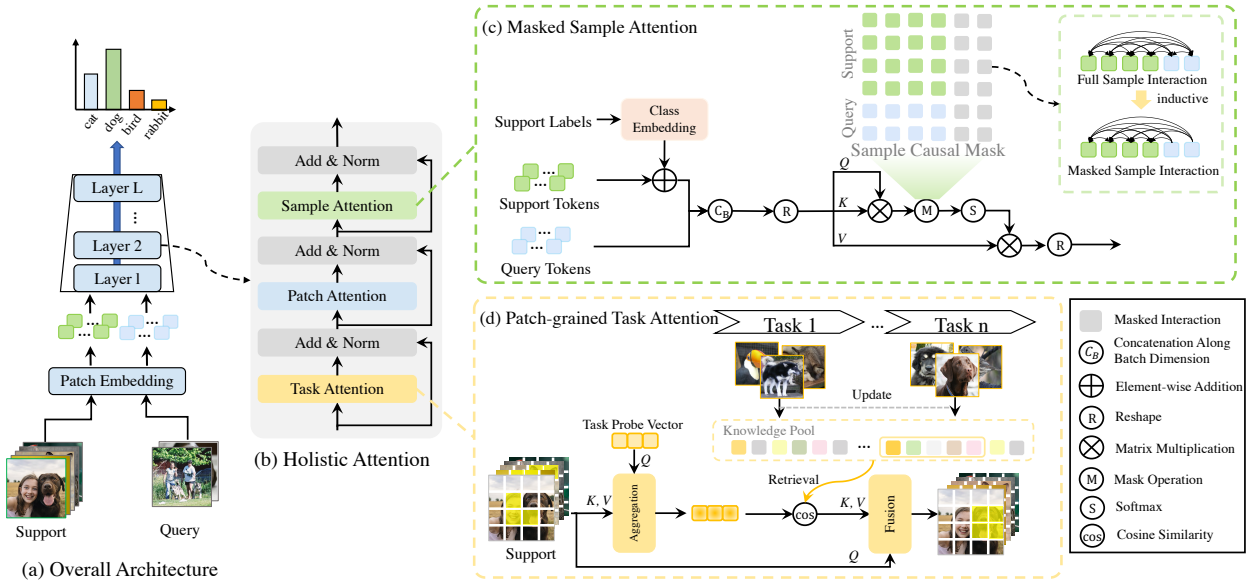


Figure 1. **Overview –** (a) The architecture of the MetaFormer with holistic attention, which extracts feature representations of support and query samples with only one feedforward while following the inductive protocol; (b) Three modules build holistic attention integrated with intra- and inter-task interaction in a sequential mode; (c) Schematic illustration of the proposed Masked Sample Attention (MSA) with sample causal attention mask to exploit the within-support and support-query relations for task-specific embeddings. (d) Schematic illustration of the proposed Patch-grained Task Attention (PTA) where foreground region is further concentrated by previous relevant semantic knowledge with the task-specific probe vector.

ficient architecture designed to decouple patch attention and sample attention, illustrated in Figure 1(b). In the case of decoupled patch-sample attention, within each layer, our approach initially computes patch-only attention as (1) to obtain features isolating backgrounds and emphasizing underlying objects. Subsequently, we reshape the token sequence to $\mathbb{R}^{L\times(NK+M)\times d}$ that is fed to MHSA with sample attention matrix $\mathbf{A}_S \in \mathbb{R}^{(NK+M)\times(NK+M)}$, incorporating sample interactions across all patches at the same patch location to capture the similarities and variances among samples, which is essential for the feature extraction in a given task to discern task-specific discriminative regions. As such, the computation complexity is reduced to $O(L(NK+M)^2+(NK+M)L^2)$. See Figure 5 in Appendix J for an illustration. Though this decoupling shares the spirit with video transformers (Ho et al., 2019; Bertasius et al., 2021), it is crucial to highlight that our consideration of the sample-to-sample relationship in few-shot learning presents a unique challenge distinct from the frame-to-frame relationship in videos, i.e., query samples have to be differentiated from support ones. This challenge motivates the following introduction of sample causal masks.

As shown in Figure 1(c), we introduce our Masked Sample Attention (MSA) with label infusion and the designed causal masking mechanism to further enforce consistency in the attended features across samples within a task. We first get the embedded support category information $\mathbf{W}^c y \in \mathbb{R}^{1\times d}$ via the linear projection matrix $\mathbf{W}^c$, which is infused to support

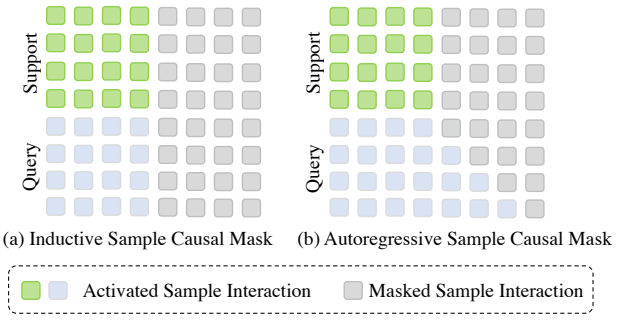


Figure 2. Sample Causal Attenntion Mask. An example with $N=4$ and $K=1$. (a) Inductive sample causal mask for within-support and support-query sample correspondence learning. (b) Autoregressive sample causal mask for extra query-query sample correspondence learning.

tokens through the elementwise addition. For the obtained sample attention $\mathbf{A}_S$, we maintain the sample causal mask $\mathbf{H} \in \mathbb{R}^{(NK+M)\times(NK+M)}$ to restrict the sample interaction patterns as:

$$\hat{\mathbf{A}}_S = \mathbf{A}_S \odot \mathbf{H} \tag{2}$$

where $\odot$ is the element-wise product.

Through the constraint, support samples can attend themselves to strengthen intra- and inter-class discriminative clues, which query samples utilize for task-specific feature consistency learning. Note that this mask mechanism also makes our method comply with the inductive protocols. In the autoregressive scenario, we also extend our MSA with the autoregressive causally-masked sample attention

to embed the query-query interactions into the vision transformer. Figure 2(b) shows an example mask with $N = 4$ and $K = 1$. Query samples attend to support samples and earlier predicted queries in an autoregressive fashion, which thus serves to implicitly expand the support set for subsequent query predictions.

3.3. Patch-grained Task Attention

In this section, we introduce details of the proposed Patch-grained Task Attention (PTA), as illustrated in Figure 1(d). PTA aims to transfer previous task knowledge for regularizing the adaptation in new tasks. To this end, PTA learns a task-specific probe vector summarizing discriminative patch-level patterns and maintains a knowledge pool consolidated during meta-training to organize learned knowledge. When a new task comes, it leverages attention-based mechanisms to tap into and aggregate relevant knowledge from the knowledge pool. We elaborate on four key components as follows: task probe vector aggregation, knowledge retrieval, pool consolidation, and knowledge fusion.

Task Probe Vector Aggregation. Given a task consisting of support and query sets, we first gather the task information with learnable task probe vectors $\mathbf{g} \in \mathbb{R}^{T\times d}$, which are computed along with support patch tokens $\mathbf{X}_{\mathcal{C}} \in \mathbb{R}^{NK\times L\times d}$ to aggregate the key parts of samples and the whole task representations. Specifically, we perform the task aggregation using attention as:

$$\mathbf{g}' = \text{Aggregation} = \text{MHSA}\left(\mathbf{Q}_{\mathbf{g}}, \mathbf{K}_{\mathbf{X}_{\mathcal{C}}}, \mathbf{V}_{\mathbf{X}_{\mathcal{C}}}\right) \quad (3)$$

where $\mathbf{Q}_{\mathbf{g}}$ is query embedding of task probe vectors; $\mathbf{K}_{\mathbf{X}_{\mathcal{C}}}$ and $\mathbf{V}_{\mathbf{X}_{\mathcal{C}}}$ are key and value embeddings of support patch tokens, respectively. This allows task probe vectors to focus on relevant task-specific feature patterns and ignore irrelevant semantics of each sample.

Knowledge Retrieval. After gathering the task information, we retrieve relevant knowledge using a simple weighted summation strategy from the knowledge pool $\mathbf{p} \in \mathbb{R}^{Z\times d}$ with Z components (which will be introduced below). The retrieval is formulated as:

$$\mathbf{r} = \sum_{z} \gamma\left(\mathbf{g}', \mathbf{p}_z\right)\mathbf{p}_z \quad (4)$$

where γ is the score function based on cosine similarity between task probe vectors and pool components. $\mathbf{r} \in \mathbb{R}^{T\times d}$ is the retrieved historical knowledge that can be thought of as key feature semantics (e.g., ears and eyes of the dog) related to the current task samples.

Pool Consolidation. During meta-training, we maintain a knowledge pool $\mathbf{p}$ updated by every sequentially coming task. To consolidate the learned knowledge in the pool, we select relevant components from the pool and integrate them with new information brought by the current task as follows:

$$\mathbf{p}_s = \mathbf{p}_s + \mathbf{g}'_j \quad (5)$$

where $s = \text{argmax}\,\gamma\left(\mathbf{g}'_j, \mathbf{p}\right)$ representing the indices of the component most similar to j-th task probe vector. This method also allows us to control the pool size and the memory consumption.

Knowledge Fusion. To regularize the adaptation of new tasks with historical knowledge, we deliver the union of original task vectors $\mathbf{g}$ and retrieved knowledge $\mathbf{r}$ to enhance the support patch token representations via the attention mechanism as follows:

$$\text{Fusion} = \text{MHSA}\left(\mathbf{Q}_{\mathbf{X}_{\mathcal{C}}}, \mathbf{K}_{[\mathbf{g}';\mathbf{r}]}, \mathbf{V}_{[\mathbf{g}';\mathbf{r}]}\right) \quad (6)$$

where $\mathbf{Q}_{\mathbf{X}_{\mathcal{C}}}$ is query embedding of support patch tokens and $\mathbf{K}_{[\mathbf{g};\mathbf{r}]}$ and $\mathbf{V}_{[\mathbf{g};\mathbf{r}]}$ are key and value embeddings of regularized task-specific semantics, respectively. The intuition is to leverage well-learned feature semantics in previous similar tasks to strengthen discriminative regions in new tasks.

3.4. Meta-training and Inference with MetaFormer

Meta-training. Using MSA and PTA as basic building blocks working in conjunction with original ViT modules, we propose a new vision transformer f_θ with holistic attention, named MetaFormer, encapsulating three distinct relationships between patch tokens, samples and tasks at different semantic levels, to extract rich task-specific feature representations. For the inductive protocol (Vinyals et al., 2016b), MetaFormer configures the sample causal mask as the inductive variant, which, coupled with the inherent Layer Normalization in vision transformers, ensures independent predictions for each query sample. Built upon feature embeddings extracted by MetaFormer, we estimate the class patch prototypes by averaging support patch tokens per class $p_k = \frac{1}{|\mathcal{S}^k|}\sum_{x\in\mathcal{S}^k} f_\theta\left(x\right)$. Query samples are predicted based on patch-wise cosine similarity with prototypes (Lai et al., 2022; Hiller et al., 2022). The probability of k^{th} category is:

$$P(\hat{y}_t = k \mid x_t) = \frac{e^{d(f_\theta(x_t), p_k)/\tau}}{\sum_c e^{d(f_\theta(x_t), p_c)/\tau}} \quad (7)$$

where d indicates the cosine distance and τ is scaling temperature. The cross-entropy loss function with the few-shot label y_t is:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{M} \log P(\hat{y}_t = y_t \mid x_t) \quad (8)$$

We also introduce the autoregressive setting from regression tasks (Nguyen & Grover, 2022; Bruinsma et al., 2023) to classification scenarios, seamlessly switching with an

autoregressive sample causal mask to allow interactions between subsequent query samples and those predicted earlier in a single pass. The support prototypes are then enriched by feeding previously predicted queries as the auxiliary support set $\tilde{\mathcal{S}}$ with predicted probability belonging to class k. We employ $P(\hat{y}_t = k \mid x_t)$ as the sample weights to compute auxiliary prototypes through a weighted average $\hat{p}_k = \frac{1}{\sum_{x \in \tilde{\mathcal{S}}^k} P(k|x)} \sum_{x \in \tilde{\mathcal{S}}^k} P(k \mid x) f_\theta(x)$. New prototypes are updated by the mean of p_k and $\hat{p}_k$. Given modeling dependencies between all M query samples requires M prototype updates, we propose managing updates with a sampling size of r queries at a time for faster and more consistent prototype refinement.

Inference. MetaFormer extracts support and query feature embeddings in a single feedforward pass during inference, configuring the inductive and autoregressive sample causal masks for inductive and autoregressive settings, respectively. For autoregressive inference, we adopt an inference-time augmentation similar to prior works (Bendou et al., 2022; Zhu & Koniusz, 2023) while differing by shuffling the order of the samples in a meta-testing task s times and then computing the most confident logits as the final prediction.

4. Experiments

4.1. Standard Few-Shot Learning

Datasets. We train and evaluate our MetaFormer on the four standard few-shot benchmarks: *mini*ImageNet (Vinyals et al., 2016b), *tiered*ImageNet (Ren et al., 2018b), CIFAR-FS (Bertinetto et al., 2019) and FC-100 (Oreshkin et al., 2018). In all experiments, we follow the standard data usage specifications same as Hiller et al. (2022), splitting data into the meta-training set, meta-validation set, and meta-test set, and classes in each set are mutually exclusive. The details of each dataset are described in Appendix A.1.

Implementation Details. We train our method in two stages following Hiller et al. (2022): self-supvervised pretraining and meta-tuning. We first pre-train our vision transformer backbone (Dosovitskiy et al., 2020; Liu et al., 2021) utilizing a self-supervised training objective (Zhou et al., 2022a). Subsequently, we integrate our proposed MSA and PTA into the vision transformer for meta-learning. We denote our framework as MetaFormer-I in the inductive setting and MetaFormer-A in the autoregressive setting. See Appendix A.2 for more training and evaluation details.

Comparison with the State-of-the-art. The comparison results with related or recent state-of-the-art (SOTA) methods on *mini*ImageNet and *tiered*ImageNet is shown in Table 1. Our method significantly outperforms previous SOTA meta-learning approaches. For instance, on *mini*ImageNet, MetaFormer-I exceeds FewTURE (Hiller et al., 2022) by $\mathbf{7.76}\%$ and $\mathbf{5.51}\%$ in 1-shot and 5-shot settings, respectively, showcasing the remarkable effectiveness of our holistic attention mechanism in harnessing the full potential of transformers for meta-learning. Additionally, MetaFormer works synergistically with self-distillation training methods (He et al., 2022b; Lin et al., 2023) to enhance task-specific feature embedding and overall performance. Results on CIFAR-FS and FC100, presented in Table 2, further validate the superiority of MetaFormer-I. As illustrated in Table 3, MetaFormer demonstrates computational efficiency over FewTURE and SMKD, primarily by eliminating the need for inference-time tuning and reducing the large resolution requirements. It is noteworthy that MetaFormer facilitates full sample interactions, as opposed to FewTURE, which focuses solely on contextual relationships. Detailed evaluations of computational cost are provided in Appendix H. Seamless transitioning from the inductive setting with the autoregressive sample causal mask, MetaFormer-A outperforms protoLP (Zhu & Koniusz, 2023) in the majority of cases by a clear margin and stands out as the first pure transformer-backed method for transductive few-shot image classification. For more comprehensive comparison results and discussions, refer to Appendix A.3.

4.2. Broader Study of Few-Shot Learning

Cross-Domain and Multi-Domain Few-shot Classification. To further investigate the fast adaptation ability of our method, we evaluate the MetaFormer in more challenging cross-domain (Chen et al., 2019; Oh et al., 2022) and multi-domain (Triantafillou et al., 2020) scenarios, containing dual category and domain shifts. Details on benchmarks and implementation are available in Appendix B and Appendix C. We evaluate MetaFomer, meta-trained on *mini*ImageNet, on cross-domain few-shot classification benchmarks in Table 4 (See more results and discussions in Appendix B.3). MetaFormer demonstrates remarkable task adaptability, surpassing previous in-domain SOTA meta-learning methods FewTURE (Hiller et al., 2022) and PMF (Hu et al., 2022) with a significant improvement of up to $\mathbf{8.77}\%$ and $\mathbf{16.86}\%$, highlighting its effectiveness in bridging domain gaps. MetaFormer also improves the transfer learning method SMKD (Lin et al., 2023) by up to $\mathbf{2.57}\%$. Furthermore, in Table 14 (Appendix C.3), we assess MetaFormer on the large-scale and challenging Meta-Dataset (Triantafillou et al., 2020), where it outperforms PMF (Hu et al., 2022) in handling tasks with diverse distributions. We attribute such impressive improvement to our proposed holistic attention mechanism, which not only facilitates sample correspondence learning but also enables knowledge reuse, thus aiding task adaptation to obtain more discriminative feature representations for each task.

Compatibility with Other Backbones and Foundation Models. Different from isotropic vision transformers, hierarchical models (Liu et al., 2021) bring greater efficiency

Table 1. **Average classification accuracy (%) for 5-way 1-shot and 5-way 5-shot scenarios.** Reported are the mean and 95% confidence interval on the unseen test sets of *mini*ImageNet (Vinyals et al., 2016a) and *tiered*ImageNet (Ren et al., 2018a). * denotes results reported by us. More comprehensive results are shown in the appendix.

Method	Setting	Backbone	# Params	*mini*ImageNet		*tiered*ImageNet	
				1-shot	5-shot	1-shot	5-shot
FEAT (Ye et al., 2020)	Inductive	*ResNet-12*	14.1 M	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16
DeepEMD (Zhang et al., 2020a)	Inductive	*ResNet-12*	12.4 M	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
RENet (Kang et al., 2021)	Inductive	*ResNet-12*	12.6 M	67.60±0.44	82.58±0.30	71.61±0.51	85.28±0.35
COSOC (Luo et al., 2021)	Inductive	*ResNet-12*	12.4 M	69.28±0.49	85.16±0.42	73.57±0.43	87.57±0.10
LEO (Rusu et al., 2018)	Inductive	*WRN-28-10*	36.8 M	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09
MetaQDA (Zhang et al., 2021c)	Inductive	*WRN-28-10*	36.5 M	67.83±0.64	84.28±0.69	74.33±0.65	89.56±0.79
SUN (Dong et al., 2022)	Inductive	*ViT*	12.5 M	67.80±0.45	83.25±0.30	72.99±0.50	86.74±0.33
FewTURE (Hiller et al., 2022)	Inductive	*ViT-Small*	22 M	68.02±0.88	84.51±0.53	72.96±0.92	86.43±0.67
PMF* (Hu et al., 2022)	Inductive	*ViT-Small*	21 M	71.01±0.81	86.03±0.53	76.61±0.89	89.66±0.54
MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**75.78±0.71**	**90.02±0.44**	**79.05±0.81**	**90.40±0.53**
HCTransformers (He et al., 2022b)	Inductive	3×*ViT-Small*	63 M	74.74±0.17	89.19±0.13	79.67±0.20	91.72±0.11
SMKD (Lin et al., 2023)	Inductive	*ViT-Small*	21 M	74.28±0.18	88.89±0.09	78.83±0.20	91.21±0.11
SMKD + MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**76.54±0.73**	**90.76±0.41**	**80.57±0.82**	**92.42±0.49**
EASY (Bendou et al., 2022)	Transductive	3×*ResNet-12*	37.2 M	84.04±0.23	89.14±0.11	84.29±0.24	89.76±0.14
protoLP (Zhu & Koniusz, 2023)	Transductive	*WRN-28-10*	36.5 M	84.32±0.21	90.02±0.12	**89.65±0.22**	93.21±0.13
MetaFormer-A (Ours)	Transductive	*ViT-Small*	24.5 M	**84.78±0.79**	**91.39±0.42**	88.38±0.78	**93.37±0.45**

Table 2. **Average classification accuracy (%) for 5-way 1-shot and 5-way 5-shot scenarios.** Reported are the mean and 95% confidence interval on the unseen test sets of CIFAR-FS (Bertinetto et al., 2019) and FC100 (Oreshkin et al., 2018). * denotes results reported by us. More comprehensive results are shown in the appendix.

Method	Setting	Backbone	# Params	CIFAR-FS		FC100	
				1-shot	5-shot	1-shot	5-shot
MetaOpt (Lee et al., 2019)	Inductive	*ResNet-12*	12.4 M	72.00±0.70	84.20±0.50	41.10±0.60	55.50±0.60
RFS (Tian et al., 2020)	Inductive	*ResNet-12*	12.4 M	73.90±0.80	86.90±0.50	44.60±0.70	60.90±0.60
BML (Zhou et al., 2021b)	Inductive	*ResNet-12*	12.4 M	73.45±0.47	88.04±0.33	45.00±0.41	63.03±0.41
TPMN (Wu et al., 2021)	Inductive	*ResNet-12*	12.4 M	75.50±0.90	87.20±0.60	46.93±0.71	63.26±0.74
PSST (Chen et al., 2021b)	Inductive	*WRN-28-10*	36.5 M	77.02±0.38	88.45±0.35	-	-
Meta-QDA (Zhang et al., 2021c)	Inductive	*WRN-28-10*	36.5 M	75.83±0.88	88.79±0.75	-	-
SUN (Dong et al., 2022)	Inductive	*ViT*	12.5M	78.37±0.46	88.84±0.32	-	-
FewTURE (Hiller et al., 2022)	Inductive	*ViT-Small*	22 M	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73
PMF* (Hu et al., 2022)	Inductive	*ViT-Small*	21 M	76.70±0.90	87.61±0.60	45.79±0.73	63.64±0.72
MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**80.16±0.76**	**90.57±0.55**	**51.14±0.71**	**68.33±0.74**
HCTransformers (He et al., 2022b)	Inductive	3×*ViT-Small*	63 M	78.89±0.18	90.50±0.09	48.27±0.15	66.42±0.16
SMKD (Lin et al., 2023)	Inductive	*ViT-Small*	21M	80.08±0.18	90.91±0.13	50.38±0.16	68.50±0.16
SMKD + MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**81.49±0.74**	**91.91±0.54**	**52.18±0.78**	**71.29±0.73**
EASY (Bendou et al., 2022)	Transductive	3×*ResNet-12*	37.2 M	87.16±0.21	90.47±0.15	54.13±0.24	66.86±0.19
protoLP (Zhu & Koniusz, 2023)	Transductive	*WRN-28-10*	36.5 M	87.69±0.23	90.82±0.15	–	–
MetaFormer-A (Ours)	Transductive	*ViT-Small*	24.5 M	**88.34±0.76**	**92.21±0.59**	**58.04±0.99**	**70.80±0.76**

Table 3. Inference efficiency comparison with existing methods on the *mini*-ImageNet.

Method	GLOPs	Inference time [ms]	
		1-shot	5-shot
FewTURE	5.01	77.35±0.47	111.22±1.27
SMKD	12.58	137.58±0.66	171.37±0.78
MetaFormer-I	4.88	**67.65±0.78**	**105.72±1.06**

and capture multi-scale information. Our evaluation of the MetaFormer on Swin-Transformer (Liu et al., 2021) (see Table15 in Appendix E) reveals consistent improvements and underscores its broad applicability. Pre-trained vision foundation models demonstrate impressive zero-shot image classification performance (Radford et al., 2021). Recent works have shown that CLIP's performance on downstream tasks can be further enhanced by utilizing few-shot data and techniques (Zhang et al., 2021b; Zhou et al., 2022b; Zhu et al., 2023). These approaches have shown promising improvements over frozen models like Zero-shot CLIP. To further investigate the adaptation ability of our proposed method, we adapt our method to CLIP model with ViT-B/16 for advancing its performance in downstream tasks. Implementation details are available in Appendix D. As shown in Table 5, CLIP pre-trained on large-scale web-crawled image-text pairs struggles with downstream datasets exhibiting a large domain gap, such as the medical dataset of ISIC. Adapting CLIP with few labeled samples is pivotal to guarantee better performance, though naively increasing the

Table 4. **Broader study of cross-domain few-shot learning.** Average classification accuracy (%) for 5-way 1-shot scenario when meta-learning on *mini*ImageNet (Vinyals et al., 2016a) but meta-testing on cross-domain few-shot benchmarks.

Method	CUB	Cars	Places	Plantae	CropDisease	EuroSAT	ISIC	ChestX
PMF (Hu et al., 2022)	40.12±0.74	33.91±0.64	60.84±0.88	43.78±0.77	72.62±0.86	64.24±0.80	31.00±0.58	22.40±0.44
FewTURE (Hiller et al., 2022)	48.21±0.83	33.97±0.63	58.74±0.91	43.31±0.76	68.22±0.88	61.77±0.81	28.67±0.56	22.60±0.44
MetaFormer-I (Ours)	**56.98±0.93**	**37.32±0.66**	**61.90±0.89**	**49.30±0.82**	**74.48±0.82**	**68.23±0.79**	**33.22±0.57**	**23.02±0.42**
SMKD (Lin et al., 2023)	54.64±0.84	34.30±0.64	62.75±0.92	45.57±0.81	75.99±0.82	68.58±0.77	33.92±0.62	22.59±0.41
SMKD+MetaFormer-I (Ours)	**57.21±0.88**	**35.38±0.64**	**62.89±0.85**	**46.13±0.80**	**76.06±0.79**	**70.04±0.82**	**34.69±0.62**	**22.65±0.40**
protoLP (Zhu & Koniusz, 2023)	**69.94±1.23**	35.50±0.93	67.44±1.31	44.26±1.15	87.13±1.10	77.89±1.20	33.00±0.78	21.70±0.44
MetaFormer-A (Ours)	67.39±0.98	**35.68±0.74**	**70.78±1.05**	**50.51±0.98**	**87.36±0.84**	**79.44±0.87**	**36.38±0.72**	**22.91±0.41**

number of parameters to adapt even incurs overfitting. Our method significantly enhances Zero-shot CLIP on EuroSAT by $\mathbf{42.76}\%$ and ISIC by $\mathbf{43.81}\%$, and its adaptation ability also surpasses Tip-Adapter by a large margin.

Table 5. **Classification results with foundation model for 1-shot EuroSAT and ISIC**. Reported are the mean and 95% confidence interval on the test set. ViT-B/16 with the patch size 16×16 is adopted for the vision branch in all methods.

Method	EuroSAT	ISIC
Zero-shot CLIP	48.73±0.98	21.07±0.76
Tip-Adapter	69.85±0.75	28.70±0.97
Tip-Adapter-F	72.01±0.97	32.27±1.11
Tip-Adapter-F with more layers	51.95±0.86	16.17±0.78
Tip-Adapter-F+MetaFormer-I (Ours)	**88.83±0.78**	**45.96±1.44**

4.3. Ablation study

Component Analysis. In this section, we investigate the individual contributions of key components in MetaFormer. The impact on performance, along with the associated increase in the number of additional learnable parameters, are detailed in Table 6, demonstrating that both components bolster performance with only a modest increase in computational overhead. We address the potential concern that observed performance gains could be attributed solely to an increase in parameters in Appendix H.3 Table 19, where we present a detailed parameter comparison to underscore that the remarkable enhancements stem from the components' inherent design. The effectiveness of the proposed adaptive sample causal masks is further evaluated in Appendix F Table 16, revealing that a lack of effective constraints between support and query samples (see Appendix H.3 Figure 4 for details of the ablated masks of within-support and support-query) results in suboptimal performance. We demonstrate in Appendix G Table 17 that PTA exhibits superior design and performance.

Table 6. Component ablation studies and the number of additional learnable parameters on *mini*ImageNet.

MSA	PTA	Add. Params.	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot
✓	✓	+3.57M	75.78 ± 0.71	90.02 ± 0.44
✓	✗	+2.01M	74.64 ± 0.76	87.53 ± 0.47
✗	✓	+1.56M	73.63 ± 0.75	87.76 ± 0.52

Decoupled Mechanism. The self-attention in traditional ViTs allows each patch token to interact with all other patch locations, especially in deeper layers. Thus, each patch token processes and stores non-local information, which is the key for vision transformers to handle potential spatial misalignment between different images. These intuitions are supported by the results we have obtained in Table 7. We here resize the image resolution to 192×192 due to the higher memory footprint of the joint attention. The results demonstrate our decoupled approximation achieves comparable performance while offering a significant reduction in computational costs. This finding supports our hypothesis that the decoupled attention mechanism can effectively leverage cross-sample information, even in the presence of misalignment between images. We conclude that our decoupled patch-sample attention achieves a trade-off between accuracy and efficiency for capturing task-specific sample interactions.

Table 7. Comparison results of joint and decoupled sample-patch attention mechanisms on *mini*ImageNet.

Method	mini 1-shot
Joint Sample-Patch Attention	73.35 ± 0.75
Decoupled Sample-Patch Attention	73.21 ± 0.76

Distance Metrics. We extend our analysis by incorporating a comparative study with the regularized Mahalanobis distance (Bateni et al., 2022) as a representative of Euclidean-based metrics, alongside our initial approach using cosine similarity. In Table 8, we observe that temperature scaling plays a pivotal role in modulating the interaction with the softmax, and scaled cosine similarity performs at par with the Euclidean distance, which aligns with the effects of metric scaling (Oreshkin et al., 2018). Furthermore, our analysis shows the resilience of cosine similarity against the intervention of noisy patch information for computing local pairwise distances.

Class Embeddings. Unlike previous meta-learning methods using class embeddings for better modeling the task-specific distribution (Rusu et al., 2018; Xu et al., 2020; Zhmoginov et al., 2022), our approach incorporates sample label information into the feature extraction process to capture nuanced distinctions between classes. The results

Table 8. Comparison results of different metrics on *mini*ImageNet. Temp.: Temperature Scaling, Patch.: Patch-wise distance.

Method	Temp.	Patch.	mini 1-shot
Cosine Distance	✗	✗	64.22 ± 0.83
	✓	✗	74.14 ± 0.73
	✓	✓	75.78 ± 0.71
Mahalanobis Distance	✗	✗	72.93 ± 0.74
	✓	✗	73.09 ± 0.75
	✓	✓	71.97 ± 0.76

in Table 9 underscore that introducing class embeddings via either concatenation or summation facilitates differentiation between classes and further strengthens task-specific discriminative clues with the guidance of a sample causal mask.

Table 9. Different label infusion methods with ViT-Small on *mini*ImageNet.

Method	Backbone	mini 1-shot
w/o. label	*ViT-Small*	73.70
concatenation	*ViT-Small*	74.59
summation	*ViT-Small*	74.64

Inference-time Augmentation Strategies. In transductive few-shot classification tasks, our model MetaFormer-A incorporates an inference-time augmentation strategy by reshuffling the sample order, while prior state-of-the-art methods involve evaluating multiple random crops per sample and subsequently averaging the features for the final prediction (Bendou et al., 2022; Qi et al., 2021; Zhu & Koniusz, 2023). We evaluate the impact of different augmentation strategies in Table 10. The results showcase the unique and synergistic role of reshuffling in enhancing the autoregressive and sequential nature of our MetaFormer-A.

Table 10. Comparison of different inference-time augmentation strategies. We report 1-shot accuracy on CIFAR-FS. Crop.: Random Cropping

Method	Backbone	Crop.	Reshuffling	1-shot
protoLP	*WRN-28-10*	✓	✗	87.69 ± 0.23
protoLP	*WRN-28-10*	✓	✓	87.69 ± 0.23
MetaFormer-A	*ViT-Small*	✗	✗	85.76 ± 0.77
MetaFormer-A	*ViT-Small*	✗	✓	88.34 ± 0.76

4.4. Qualitative Analysis

Figure 3 presents visualizations of our holistic attention mechanism, with columns depicting the attention maps of three distinct attention modules. The results reveal that the sample correspondence learning guided by spatial and sample attention modules suppresses irrelevant regions by exploiting pattern relations within and across samples, leading to the extraction of more discriminative, task-specific features. Furthermore, the task attention module adeptly transfers semantic knowledge from previous tasks to new ones, with a particular focus on the critical components of foreground objects. When integrated with intra- and inter-task attention, holistic attention yields a more precise and comprehensive response map, predominantly concentrated on the foreground region.

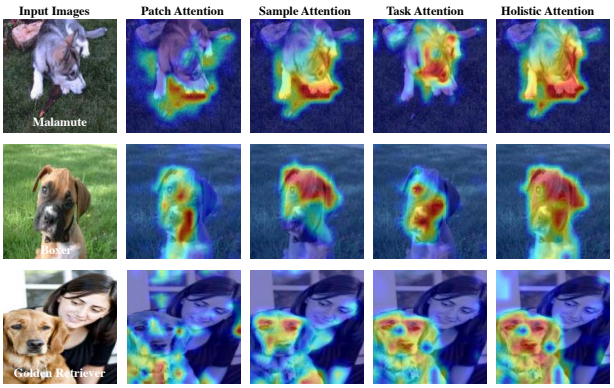


Figure 3. Response visualization for MetaFormer with holistic attention.

5. Conclusions

This paper proposes MetaFormer, a novel Vit-backed meta-learning approach for few-shot classification. MetaFormer capitalizes on the transformer architecture to orchestrate holistic attention, integrating two lightweight modules to capture intra-task and inter-task interactions. Through Masked Sample Attention (MSA), MetaFormer promotes sample consistency for adapting task-specific discriminative feature representations. Meanwhile, Patch-grained Task Attention (PTA) leverages a dynamic knowledge pool to infuse relevant historical knowledge into current task adaptation. Configuring different adaptive masks, MetaFormer supports both inductive and transductive few-shot learning protocols. Extensive experiments demonstrate the superiority of MetaFormer across standard in-domain, cross-domain, and multi-domain benchmarks.

Impact Statement

Broader Impact. Our proposed meta-learning method integrates seamlessly with recent state-of-the-art pre-trained vision transformers and foundation models, significantly enhancing their few-shot classification capabilities, and hence has the potential to advance the field of machine learning. The deployment of such advanced machine learning models in real-world scenarios must be governed by careful consideration of privacy, fairness, and transparency to ensure responsible use.

Limitations. The current study primarily focuses on few-shot image classification, leaving the exploration of our method's applicability to detection and segmentation tasks for future work.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6430–6439, 2019.

Afrasiyabi, A., Lalonde, J.-F., and Gagné, C. Mixture-based feature space learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9041–9051, 2021.

Afrasiyabi, A., Larochelle, H., Lalonde, J.-F., and Gagné, C. Matching feature sets for few-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9014–9024, June 2022.

Bao, H., Dong, L., Piao, S., and Wei, F. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Bateni, P., Goyal, R., Masrani, V., Wood, F., and Sigal, L. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14493–14502, 2020.

Bateni, P., Barber, J., Goyal, R., Masrani, V., van de Meent, J.-W., Sigal, L., and Wood, F. Beyond simple meta-learning: Multi-purpose models for multi-domain, active and continual few-shot learning. *arXiv preprint arXiv:2201.05151*, 2022.

Bendou, Y., Hu, Y., Lafargue, R., Lioi, G., Pasdeloup, B., Pateux, S., and Gripon, V. Easy—ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components. *Journal of Imaging*, 8 (7):179, 2022.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Bertinetto, L., Henriques, J. F., Torr, P., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Bruinsma, W. P., Markou, S., Requeima, J., Foong, A. Y. K., Andersson, T. R., Vaughan, A., Buonomo, A., Hosking, J. S., and Turner, R. E. Autoregressive conditional neural processes. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C., and Huang, J.-B. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Chen, Y., Liu, Z., Xu, H., Darrell, T., and Wang, X. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9062–9071, 2021a.

Chen, Z., Ge, J., Zhan, H., Huang, S., and Wang, D. Pareto self-supervised training for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13663–13672, 2021b.

Doersch, C., Gupta, A., and Zisserman, A. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993, 2020.

Dong, B., Zhou, P., Yan, S., and Zuo, W. Self-promoted supervision for few-shot transformer. In *European Conference on Computer Vision*, pp. 329–347, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Douillard, A., Ramé, A., Couairon, G., and Cord, M. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9285–9295, 2022.

Fei, N., Lu, Z., Xiang, T., and Huang, S. Melr: Meta-learning via modeling episode-level relationships for few-shot learning. In *International Conference on Learning Representations*, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135, 2017.

Gao, Z., Wu, Y., Jia, Y., and Harandi, M. Curvature generation in curved spaces for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8691–8700, 2021.

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., and Cord, M. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8059–8068, 2019.

Guo, Y., Codella, N. C., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., and Feris, R. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pp. 124–141, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022a.

He, Y., Liang, W., Zhao, D., Zhou, H.-Y., Ge, W., Yu, Y., and Zhang, W. 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*, pp. 9119–9129, 2022b.

Helber, P., Bischke, B., Dengel, A., and Borth, D. 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.

Hiller, M., Ma, R., Harandi, M., and Drummond, T. Rethinking generalization in few-shot classification. *arXiv preprint arXiv:2206.07267*, 2022.

Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Hou, R., Chang, H., Ma, B., Shan, S., and Chen, X. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Hu, S. X., Li, D., Stühmer, J., Kim, M., and Hospedales, T. M. 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*, pp. 9068–9077, 2022.

Hu, Y., Gripon, V., and Pateux, S. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks*, pp. 487–499, 2021.

Jiang, W., Kwok, J., and Zhang, Y. Subspace learning for effective meta-learning. In *International Conference on Machine Learning*, pp. 10177–10194, 2022.

Kang, D., Kwon, H., Min, J., and Cho, M. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Kim, J., Kim, H., and Kim, G. Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *European Conference on Computer Vision*, pp. 599–617, 2020.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Lai, J., Yang, S., Liu, W., Zeng, Y., Huang, Z., Wu, W., Liu, J., Gao, B.-B., and Wang, C. tsf: Transformer-based semantic filter for few-shot learning. In *European Conference on Computer Vision*, pp. 1–19, 2022.

Lazarou, M., Stathaki, T., and Avrithis, Y. Iterative label cleaning for transductive and semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8751–8760, 2021.

Lee, K., Maji, S., Ravichandran, A., and Soatto, S. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

Lin, H., Han, G., Ma, J., Huang, S., Lin, X., and Chang, S.-F. Supervised masked knowledge distillation for few-shot transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19649–19659, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Luo, X., Wei, L., Wen, L., Yang, J., Xie, L., Xu, Z., and Tian, Q. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Ma, J., Xie, H., Han, G., Chang, S.-F., Galstyan, A., and Abd-Almageed, W. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10573–10582, 2021a.

Ma, R., Fang, P., Avraham, G., Zuo, Y., Drummond, T., and Harandi, M. Learning instance and task-aware dynamic kernels for few shot learning. *arXiv preprint arXiv:2112.03494*, 2021b.

Ma, R., Fang, P., Drummond, T., and Harandi, M. Adaptive poincaré point to set distance for few-shot classification. *arXiv preprint arXiv:2112.01719*, 2021c.

Nguyen, T. and Grover, A. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. *arXiv preprint arXiv:2207.04179*, 2022.

Oh, J., Kim, S., Ho, N., Kim, J.-H., Song, H., and Yun, S.-Y. Understanding cross-domain few-shot learning: An experimental study. *arXiv preprint arXiv:2202.01339*, 2022.

Oreshkin, B., Rodríguez López, P., and Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Qi, G., Yu, H., Lu, Z., and Li, S. Transductive few-shot classification on the oblique manifold. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8412–8422, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

Ranftl, R., Bochkovskiy, A., and Koltun, V. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018a.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018b.

Requeima, J., Gordon, J., Bronskill, J., Nowozin, S., and Turner, R. E. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2018.

Simon, C., Koniusz, P., Nock, R., and Harandi, M. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4136–4145, 2020.

Smith, J. S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., and Kira, Z. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Thrun, S. and Pratt, L. *Learning to learn*. Springer Science & Business Media, 2012.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pp. 266–282, 2020.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357, 2021.

Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.-A., et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020.

Triantafillou, E., Larochelle, H., Zemel, R., and Dumoulin, V. Learning a universal template for few-shot dataset generalization. In *International Conference on Machine Learning*, pp. 10424–10433, 2021.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Tseng, H.-Y., Lee, H.-Y., Huang, J.-B., and Yang, M.-H. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29:3630–3638, 2016a.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29, 2016b.

Wang, Z., Liu, L., Duan, Y., Kong, Y., and Tao, D. Continual learning with lifelong vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 171–181, 2022a.

Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., and Pfister, T. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Wertheimer, D., Tang, L., and Hariharan, B. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8012–8021, 2021.

Wu, J., Zhang, T., Zhang, Y., and Wu, F. Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8433–8442, 2021.

Xie, J., Long, F., Lv, J., Wang, Q., and Li, P. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7972–7981, 2022.

Xu, C., Fu, Y., Liu, C., Wang, C., Li, J., Huang, F., Zhang, L., and Xue, X. Learning dynamic alignment via meta-filter for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5182–5191, 2021.

Xu, J., Ton, J.-F., Kim, H., Kosiorek, A., and Teh, Y. W. Metafun: Meta-learning with iterative functional updates. In *International Conference on Machine Learning*, pp. 10617–10627, 2020.

Yao, H., Wei, Y., Huang, J., and Li, Z. Hierarchically structured meta-learning. In *International Conference on Machine Learning*, pp. 7045–7054, 2019.

Yao, H., Wu, X., Tao, Z., Li, Y., Ding, B., Li, R., and Li, Z. Automated relational meta-learning. *arXiv preprint arXiv:2001.00745*, 2020.

Ye, H.-J., Hu, H., Zhan, D.-C., and Sha, F. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8808–8817, 2020.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, C., Cai, Y., Lin, G., and Shen, C. 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*, pp. 12203–12213, 2020a.

Zhang, C., Ding, H., Lin, G., Li, R., Wang, C., and Shen, C. Meta navigator: Search for a good adaptation policy for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9435–9444, 2021a.

Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Liu, X., and Liu, Z. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 7277–7286, 2022.

Zhang, M., Zhang, J., Lu, Z., Xiang, T., Ding, M., and Huang, S. Iept: Instance-level and episode-level pretext tasks for few-shot learning. In *International Conference on Learning Representations*, 2020b.

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021b.

Zhang, X., Meng, D., Gouk, H., and Hospedales, T. M. Shallow bayesian meta learning for real-world few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 651–660, 2021c.

Zhao, J., Yang, Y., Lin, X., Yang, J., and He, L. Looking wider for better adaptive representation in few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10981–10989, 2021.

Zhmoginov, A., Sandler, M., and Vladymyrov, M. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. In *International Conference on Machine Learning*, pp. 27075–27098, 2022.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022b.

Zhou, P., Zou, Y., Yuan, X.-T., Feng, J., Xiong, C., and Hoi, S. Task similarity aware meta learning: Theory-inspired improvement on maml. In *Uncertainty in artificial intelligence*, pp. 23–33, 2021a.

Zhou, Z., Qiu, X., Xie, J., Wu, J., and Zhang, C. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8402–8411, 2021b.

Zhu, H. and Koniusz, P. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23996–24006, 2023.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Zhu, X., Zhang, R., He, B., Zhou, A., Wang, D., Zhao, B., and Gao, P. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. *arXiv preprint arXiv:2304.01195*, 2023.

One Meta-tuned Transformer is What You Need for Few-shot Learning
-Supplementary Material-

A. Setup for In-Domain Few-Shot Evaluation

A.1. Datasets Used for Benchmarks

For standard few-shot image classification evaluation with only class shift, we train and evaluate our MetaFormer presented in the main paper on the following few-shot benchmarks: ***mini*ImageNet.** (Vinyals et al., 2016b) is a subset of the ImageNet-1K, consisting of 100 classes and 600 images in each category. The classes are divided into 64, 16, and 20 for training, validation, and test, respectively. ***tiered*ImageNet.** (Ren et al., 2018b) is another larger and more challenging subset of ImageNet-1K. It contains 34 higher-level nodes near the root of ImageNet, which are 608 classes in total. The dataset is split into 20, 6, and 8 higher-level nodes and corresponding 351, 97, and 160 classes as the training, validation, and testing set, respectively. **CIFAR-FS** (Bertinetto et al., 2019) contains 100 classes and 600 images from the CIFAR100 dataset (Krizhevsky & Hinton, 2009). The classes are split into 64 for training, 16 for validation, and 20 for testing. **FC100** (Oreshkin et al., 2018) is built from the CIFAR100 (Krizhevsky & Hinton, 2009) employing a splitting strategy analogous to that of the *tiered*ImageNet dataset to enhance difficulty, giving rise to 60 training, 20 validation, and 20 test classes.

A.2. Additional Implementation Details

Pretraining. For MetaFormer, we adhere to the strategy delineated by Hiller et al. (2022) for pretraining our vision transformer backbones on the meta-training split of each dataset, maintaining most of the training hyperparameter configurations reported in their study. Concretely, we employ default two global crops and ten local crops with respective crop scales of (0.4, 1.0) and (0.05, 0.4). We use the image resolution of 224×224 and the output is projected to 8192 dimensions. A patch size of 16 and window size of 7 are used for aligning standard settings in ViT-small (Dosovitskiy et al., 2020; Touvron et al., 2021) and Swin-tiny (Liu et al., 2021), respectively. A batch size of 512 and a cosine-decaying learning rate schedule are used. For SMKD-MetaFormer, we follow Lin et al. (2023) to train the vision transformer backbones with the image resolution of 480×480.

Meta-tuning. We integrate our proposed MSA and PTA into the original vision transformer in every other layer, starting from the 6th layer, to construct holistic attention for meta-learning. Here, MSA and PTA are randomly initialized. The number of task probe vectors T is configured to one for ViT and eight for Swin, and the pool size is set to $Z = 50$. As training progresses, we consolidate via Eqn. (5) for a new task which is sufficiently similar to previous components, where we evaluate the similarity via cosine similarity and set a similarity threshold of 0.5; otherwise, we directly add the task-specific g' to the pool if there is available capacity. During meta-tuning, we follow most of the training techniques used in FewTURE (Hiller et al., 2022). We employ the SGD optimizer, utilizing a cosine-decaying learning rate initiated at 2×10^{-4}, a momentum value of 0.9, and a weight decay of 5×10^{-4} across all datasets. The input image size is set to 224×224 for MetaFormer and 360×360 for SMKD-MetaFormer. Typically, training is conducted for a maximum of 200 epochs. To mitigate the risk of overfitting, we adopt the early stopping strategy coupled with freezing parameters of the first three layers. For a fair comparison with the state-of-the-art transductive methods (Lazarou et al., 2021; Zhu & Koniusz, 2023), we adopt the same feature pre-processing as Hu et al. (2021) for MetaFormer-A. We set the sampling size $r = 15$ and shuffle the order of samples $s = 30$ times for autoregressive inference. All additional hyperparameters are selected on 600 randomly sampled episodes from the respective validation sets to ascertain the optimal parameter configuration. For the evaluation of few-shot learning, we conduct a random sampling of 600 episodes from the test set to evaluate our model.

A.3. Results

We present a more comprehensive few-shot evaluation in Table 11 and Table 12, comparing various methods on the four benchmark datasets. Based on ViT-Small, our MetaFormer consistently outperforms established meta-learning methods PMF (Hu et al., 2022) and FewTURE (Hiller et al., 2022) across all evaluations. Notably, this remarkable performance is achieved without the need for inference-time fine-tuning, underscoring its robustness and effectiveness. We find that our MetaFormer effectively synergizes with transfer learning method SMKD (Lin et al., 2023) to further enhance its performance. When compared to the transductive protoLP (Zhu & Koniusz, 2023), MetaFormer exhibits superior performance in most scenarios with a smaller number of parameters.

Table 11. **More comprehensive results on *mini*ImageNet and *tiered*ImageNet for 5-way 1-shot and 5-way 5-shot scenarios.** Reported are the mean and 95% confidence interval of average classification accuracy (%) on the unseen meta-test set. * denotes results reported by us.

Method	Setting	Backbone	# Params	*mini*ImageNet		*tiered*ImageNet	
				1-shot	5-shot	1-shot	5-shot
MatchNet (Vinyals et al., 2016b)	Inductive	*ResNet-12*	12.4 M	61.24±0.29	73.93±0.23	71.01±0.33	83.12±0.24
ProtoNet (Snell et al., 2017)	Inductive	*ResNet-12*	12.4 M	62.29±0.33	79.46±0.48	68.25±0.23	84.01±0.56
FEAT (Ye et al., 2020)	Inductive	*ResNet-12*	14.1 M	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16
DeepEMD (Zhang et al., 2020a)	Inductive	*ResNet-12*	12.4 M	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
IEPT (Zhang et al., 2020b)	Inductive	*ResNet-12*	12.4 M	67.05±0.44	82.90±0.30	72.24±0.50	86.73±0.34
MELR (Fei et al., 2020)	Inductive	*ResNet-12*	14.1 M	67.40±0.43	83.40±0.28	72.14±0.51	87.01±0.35
FRN (Wertheimer et al., 2021)	Inductive	*ResNet-12*	12.4 M	66.45±0.19	82.83±0.13	72.06±0.22	86.89±0.14
CG (Zhao et al., 2021)	Inductive	*ResNet-12*	12.4 M	67.02±0.20	82.32±0.14	71.66±0.23	85.50±0.15
DMF (Xu et al., 2021)	Inductive	*ResNet-12*	12.4 M	67.76±0.46	82.71±0.31	71.89±0.52	85.96±0.35
BML (Zhou et al., 2021b)	Inductive	*ResNet-12*	12.4 M	67.04±0.63	83.63±0.29	68.99±0.50	85.49±0.34
CNL (Zhao et al., 2021)	Inductive	*ResNet-12*	12.4 M	67.96±0.98	83.36±0.51	73.42±0.95	87.72±0.75
Meta-NVG (Zhang et al., 2021a)	Inductive	*ResNet-12*	12.4 M	67.14±0.80	83.82±0.51	74.58±0.88	86.73±0.61
RENet (Kang et al., 2021)	Inductive	*ResNet-12*	12.6 M	67.60±0.44	82.58±0.30	71.61±0.51	85.28±0.35
PAL (Ma et al., 2021a)	Inductive	*ResNet-12*	12.4 M	69.37±0.64	84.40±0.44	72.25±0.72	86.95±0.47
COSOC (Luo et al., 2021)	Inductive	*ResNet-12*	12.4 M	69.28±0.49	85.16±0.42	73.57±0.43	87.57±0.10
Meta DeepBDC (Xie et al., 2022)	Inductive	*ResNet-12*	12.4 M	67.34±0.43	84.46±0.28	72.34±0.49	87.31±0.32
LEO (Rusu et al., 2018)	Inductive	*WRN-28-10*	36.8 M	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09
MetaFun (Xu et al., 2020)	Inductive	*WRN-28-10*	37.7 M	62.12±0.30	78.20±0.16	67.72±0.14	83.28±0.12
CC+rot (Gidaris et al., 2019)	Inductive	*WRN-28-10*	36.5 M	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36
FEAT (Ye et al., 2020)	Inductive	*WRN-28-10*	38.1 M	65.10±0.20	81.11±0.14	70.41±0.23	84.38±0.16
MetaQDA (Zhang et al., 2021c)	Inductive	*WRN-28-10*	36.5 M	67.83±0.64	84.28±0.69	74.33±0.65	89.56±0.79
OM (Qi et al., 2021)	Inductive	*WRN-28-10*	36.5 M	66.78±0.30	85.29±0.41	71.54±0.29	87.79±0.46
SUN (Dong et al., 2022)	Inductive	*ViT*	12.5 M	67.80±0.45	83.25±0.30	72.99±0.50	86.74±0.33
FewTURE (Hiller et al., 2022)	Inductive	*ViT-Small*	22 M	68.02±0.88	84.51±0.53	72.96±0.92	86.43±0.67
FewTURE (Hiller et al., 2022)	Inductive	*Swin-Tiny*	29 M	72.40±0.78	86.38±0.49	76.32±0.87	89.96±0.55
PMF* (Hu et al., 2022)	Inductive	*ViT-Small*	21 M	71.01±0.81	86.03±0.53	76.61±0.89	89.66±0.54
MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**75.78±0.71**	**90.02±0.44**	**79.05±0.81**	**90.40±0.53**
HCTransformers (He et al., 2022b)	Inductive	3×*ViT-Small*	63 M	74.74±0.17	89.19±0.13	79.67±0.20	91.72±0.11
SMKD (Lin et al., 2023)	Inductive	*ViT-Small*	21 M	74.28±0.18	88.89±0.09	78.83±0.20	91.21±0.11
SMKD + MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**76.54±0.73**	**90.76±0.41**	**80.57±0.82**	**92.42±0.49**
PT+MAP (Hu et al., 2021)	Transductive	*WRN-28-10*	36.5 M	82.92±0.26	88.82±0.13	–	–
iLPC (Lazarou et al., 2021)	Transductive	*WRN-28-10*	36.5 M	83.05±0.79	88.82±0.42	88.50±0.75	92.46±0.42
EASY (Bendou et al., 2022)	Transductive	3×*ResNet-12*	37.2 M	84.04±0.23	89.14±0.11	84.29±0.24	89.76±0.14
protoLP (Zhu & Koniusz, 2023)	Transductive	*WRN-28-10*	36.5 M	84.32±0.21	90.02±0.12	**89.65±0.22**	93.21±0.13
MetaFormer-A (Ours)	Transductive	*ViT-Small*	24.5 M	**84.78±0.79**	**91.39±0.42**	88.38±0.78	**93.37±0.45**

B. Setup for Cross-Domain Few-Shot Evaluation

B.1. Datasets Used for Benchmarks

We use *mini*ImageNet as the source dataset for meta-training and perform the cross-domain few-shot evaluation on eight datasets with varying domain similarity, following Oh et al. (2022). The datasets can be separated into two groups: **BSCD-FSL benchmark** (Guo et al., 2020) and **nonBSCD-FSL**. For BSCD-FSL benchmark (CropDisease, EuroSAT, ISIC, ChestX), we follow Guo et al. (2020) for the dataset split. And for nonBSCD-FSL benchmark (CUB, Car, Plantaem Places), we follow Tseng et al. (2020) for the splitting procedure. We refer to Oh et al. (2022) for a more detailed description of each dataset.

B.2. Implementation Details

In our cross-domain experiments, we meta-train our MetaFormer-I and MetaFormer-A on the *mini*ImageNet dataset as in Section 4.1 in the main paper and then freeze all parameters during evaluation on cross-domain benchmarks. For a fair comparison, we adhere to the standard meta-testing procedure to assess the performance of baseline models trained on *mini*ImageNet, including FewTURE (Hiller et al., 2022), PMF (Hu et al., 2022), SMKD (Lin et al., 2023) and protoLP (Zhu & Koniusz, 2023).

Table 12. **More comprehensive results on CIFAR-FS and FC100 for 5-way 1-shot and 5-way 5-shot scenarios.** Reported are the mean and 95% confidence interval of average classification accuracy (%) on the unseen meta-test set. * denotes results reported by us.

Method	Setting	Backbone	# Params	CIFAR-FS		FC100	
				1-shot	5-shot	1-shot	5-shot
ProtoNet (Snell et al., 2017)	Inductive	*ResNet-12*	12.4 M	-	-	41.54±0.76	57.08±0.76
MetaOpt (Lee et al., 2019)	Inductive	*ResNet-12*	12.4 M	72.00±0.70	84.20±0.50	41.10±0.60	55.50±0.60
MABAS (Kim et al., 2020)	Inductive	*ResNet-12*	12.4 M	73.51±0.92	85.65±0.65	42.31±0.75	58.16±0.78
RFS (Tian et al., 2020)	Inductive	*ResNet-12*	12.4 M	73.90±0.80	86.90±0.50	44.60±0.70	60.90±0.60
BML (Zhou et al., 2021b)	Inductive	*ResNet-12*	12.4 M	73.45±0.47	88.04±0.33	45.00±0.41	63.03±0.41
CG (Gao et al., 2021)	Inductive	*ResNet-12*	12.4 M	73.00±0.70	85.80±0.50	-	-
Meta-NVG (Zhang et al., 2021a)	Inductive	*ResNet-12*	12.4 M	74.63±0.91	86.45±0.59	46.40±0.81	61.33±0.71
RENet (Kang et al., 2021)	Inductive	*ResNet-12*	12.6 M	74.51±0.46	86.60±0.32	-	-
TPMN (Wu et al., 2021)	Inductive	*ResNet-12*	12.4 M	75.50±0.90	87.20±0.60	46.93±0.71	63.26±0.74
MixFSL (Afrasiyabi et al., 2021)	Inductive	*ResNet-12*	12.4 M	-	-	44.89±0.63	60.70±0.60
CC+rot (Gidaris et al., 2019)	Inductive	WRN-28-10	73.62±0.31	86.05±0.22	-	-	
PSST (Chen et al., 2021b)	Inductive	*WRN-28-10*	36.5 M	77.02±0.38	88.45±0.35	-	-
Meta-QDA (Zhang et al., 2021c)	Inductive	*WRN-28-10*	36.5 M	75.83±0.88	88.79±0.75	-	-
SUN (Dong et al., 2022)	Inductive	*ViT*	12.5M	78.37±0.46	88.84±0.32	-	-
FewTURE (Hiller et al., 2022)	Inductive	*ViT-Small*	22 M	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73
FewTURE (Hiller et al., 2022)	Inductive	*Swin-Tiny*	29 M	77.76±0.81	88.90±0.59	47.68±0.78	63.81±0.75
PMF* (Hu et al., 2022)	Inductive	*ViT-Small*	21 M	76.70±0.90	87.61±0.60	45.79±0.73	63.64±0.72
MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**80.16±0.76**	**90.57±0.55**	**51.14±0.71**	**68.33±0.74**
HCTransformers (He et al., 2022b)	Inductive	3×*ViT-Small*	63 M	78.89±0.18	90.50±0.09	48.27±0.15	66.42±0.16
SMKD (Lin et al., 2023)	Inductive	*ViT-Small*	21M	80.08±0.18	90.91±0.13	50.38±0.16	68.50±0.16
SMKD + MetaFormer-I (Ours)	Inductive	*ViT-Small*	24.5 M	**81.49±0.74**	**91.91±0.54**	**52.18±0.78**	**71.29±0.73**
PT+MAP (Hu et al., 2021)	Transductive	*WRN-28-10*	36.5 M	87.69±0.23	90.68±0.15	–	–
iLPC (Lazarou et al., 2021)	Transductive	*WRN-28-10*	36.5 M	86.51±0.75	90.60±0.48	–±–	–±–
EASY (Bendou et al., 2022)	Transductive	3×*ResNet-12*	37.2 M	87.16±0.21	90.47±0.15	54.13±0.24	66.86±0.19
protoLP (Zhu & Koniusz, 2023)	Transductive	*WRN-28-10*	36.5 M	87.69±0.23	90.82±0.15	–	–
MetaFormer-A (Ours)	Transductive	*ViT-Small*	24.5 M	**88.34±0.76**	**92.21±0.59**	**58.04±0.99**	**70.80±0.76**

B.3. Results.

In Table 13, we present a comparative evaluation of MetaFormer against other previous state-of-the-art methods in the 5-way 5-shot cross-domain scenario. MetaFormer exhibits superior performance across various domains, surpassing its meta-learning counterparts in both inductive and transductive settings by significant margins of up to 8.20% and 22.60%, respectively. Notably, MetaFormer outperforms PMF (Hu et al., 2022) in most cases by a clear margin, achieving this without test-time full fine-tuning. Our results also show that MetaFormer improves the transfer learning method SMKD (Lin et al., 2023) by up to 3.70%. These findings underscore the robustness and effectiveness of MetaFormer in task adaptation.

Table 13. **Broader study of cross-domain few-shot learning.** Average classification accuracy (%) for 5-way 5-shot scenario when meta-learning on *mini*ImageNet (Vinyals et al., 2016a) but meta-testing on cross-domain few-shot benchmarks. † means test-time finetuning is employed.

Method	CUB	Cars	Places	Plantae	CropDisease	EuroSAT	ISIC	ChestX
PMF† (Hu et al., 2022)	63.42±0.74	**49.23±0.75**	77.52±0.69	63.89±0.71	89.17±0.55	84.02±0.55	44.76±0.56	25.84±0.43
FewTURE† (Hiller et al., 2022)	67.70±0.77	46.54±0.73	74.70±0.69	61.72±0.71	86.41±0.56	77.88±0.57	38.53±0.54	25.54±0.43
MetaFormer-I (Ours)	**75.90±0.77**	48.62±0.79	**78.42±0.65**	**65.40±0.74**	**89.72±0.58**	**85.01±0.51**	**45.90±0.57**	**26.47±0.45**
SMKD (Lin et al., 2023)	77.17±0.69	50.72±0.71	80.52±0.64	63.98±0.72	92.11±0.45	85.28±0.54	47.58±0.62	25.75±0.43
SMKD+MetaFormer-I (Ours)	**80.87±0.68**	**52.68±0.74**	**80.96±0.67**	**65.60±0.72**	**92.91±0.45**	**86.60±0.50**	**49.80±0.60**	**26.25±0.45**
protoLP (Zhu & Koniusz, 2023)	76.95±1.10	44.49±0.93	74.81±1.11	54.67±1.07	82.96±1.57	65.99±1.86	37.74±0.83	23.18±0.45
MetaFormer-A (Ours)	**82.92±0.68**	**52.92±0.81**	**83.18±0.62**	**67.11±0.74**	**94.66±0.44**	**88.59±0.48**	**49.55±0.62**	**25.86±0.45**

C. Setup for Multi-Domain Few-Shot Evaluation

C.1. Datasets used for Benchmarks

Meta-Dataset (Triantafillou et al., 2020) is a more challenging and realistic large-scale benchmark consisting of ten image datasets including ImageNet-1k, Omniglot, Aircraft, CUB, Textures, QuickDraw, Fungi, VGG Flower, Traffic Signs, and

MSCOCO, each with specified train, val and test splits. We follow Hu et al. (2022) to utilize the train and val splits of the initial eight datasets (in-domain) for meta-training and validation, while employing the test splits of all datasets for meta-testing. We refer to Triantafillou et al. (2020) for an in-depth exploration of Meta-Dataset.

C.2. Implementation Details

We meta-train both PMF (Hu et al., 2022) and our MetaFormer build upon the same pre-trained vision transformer (Caron et al., 2021) in a 5-way 1-shot setting, adhering to most of the unchanged training hyperparameters reported in PMF. The meta-tuning phase includes a 10-epoch warm-up, followed by a total of 100 epochs of training. We use the SGD optimizer with a momentum of 0.9 and a cosine-decaying learning rate scheduler is employed, initialized at 5×10^{-4}.

C.3. Results.

We evaluate the effectiveness of MetaFormer on the large-scale and challenging Meta-Dataset in multi-domain scenarios. Table 14 presents the test accuracy measured on each dataset meta-test set. MetaFormer achieves comparable or superior performance compared to previous state-of-the-art meta-learning method, PMF (Hu et al., 2022), in adapting to most domains. The superior overall performance of MetaFormer, especially in settings with scarce samples (e.g., one sample per category), underscores the efficacy of our proposed approach for fast adaptation in each domain. Additionally, the computational efficiency of MetaFormer-I facilitates quicker deployment, making our method particularly suitable for real-world applications in resource-constrained scenarios.

Table 14. **Broader study of multi-domain few-shot learning.** Average classification accuracy (%) for 5-way 1-shot scenario.

Model	FT	In-domain								Out-of-domain		Avg
		INet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
PMF (Hu et al., 2022)	Y	56.35	94.22	88.00	84.63	**52.90**	75.18	84.13	75.20	55.02	49.69	71.53
MetaFormer-I (Ours)	-	**63.41**	**94.57**	87.93	**89.17**	51.33	75.10	81.97	**85.06**	**57.33**	**53.64**	**73.95**
MetaFormer-I (Ours)	Y	**64.25**	**94.64**	87.93	**89.11**	56.52	75.10	81.97	**85.31**	**62.86**	**53.82**	**75.15**
MetaFormer-A (Ours)	-	**66.03**	**96.47**	**89.75**	**91.95**	52.20	**79.17**	**84.44**	**88.88**	**58.89**	**58.12**	**76.59**

D. MetaFormer on Large-Scale Foundation Models

We evaluate different methods on 1-shot EuroSAT (Helber et al., 2019) and ISIC (Tschandl et al., 2018) datasets. For both training and testing phases, we employ the episodic approach as described in Vinyals et al. (2016b). Note that we strictly follow the TiP-Adapter-F (Zhang et al., 2021b) pipeline to sample support set from the train set and query set from the test set to build the task for evaluation since there are no new classes in the test set. For instance, in the EuroSAT dataset with 10 classes, we construct the 10-way 1-shot task, where the support and query sets are drawn from the train and test split of the EuroSAT dataset, respectively. We also integrate the cache model from TiP-Adapter as the auxiliary classifier head. We only fine-tune introduced MSA and keep frozen the visual encoder and textual encoder of CLIP. We train our method for 20 epochs on both datasets and we employ the SGD optimizer with a cosine-decaying learning rate initiated at 2×10^{-4}, a momentum value of 0.9, and a weight decay of 5×10^{-4}. We test using the pre-trained word embeddings of a single prompt, "a photo of a [CLASS]." for all methods.

E. MetaFormer on Hierarchical Transformers.

We extend MetaFormer to integrate with Swin (Liu et al., 2021), which employs shifted local window for performing self-attention within each window and merges patch embeddings to build hierarchical structures, thereby aggregating multi-scale information. As shown in Table 15, the experiments are conducted on the same pre-trained Swin-Tiny model and the results consistently demonstrate that MetaFormer outperforms FewTURE in both settings. These results highlight the versatility and general applicability of our method.

F. Abaltion of Other Design Strategies

In Table 16, we explore various design choices for our approach. The variability of feature semantics across different layers in the backbone leads us to investigate which layer optimally facilitates sample interaction. The results are shown in Table 16a, indicating that starting to build intra-task interaction from stage 6 is moderate. Notably, the integration of

Table 15. **Comparison results with the Swin-Transformer backbone on *mini*Imagenet**.

Method	Backbone	1-shot	5-shot
FewTURE (Hiller et al., 2022)	*Swin-Tiny*	72.40±0.78	86.38±0.49
MetaFormer-I (Ours)	*Swin-Tiny*	**74.17**±0.73	**89.17**±0.45

Table 16. Comparison results of different architecture design strategies.

(a) Different MSA locations in ViT-Samll.

Location	mini-1s (%)
[5, 7, 9]	74.36
[6, 8, 10]	74.64
[11]	72.34

(b) Different MSA variants with ViT-Small.

Method	mini-1s (%)
within-support	73.30
support-query	73.28
global features	71.49

(c) The number of probe vectors with ViT-Small.

Number	mini-1s (%)
1	75.78
4	75.28
8	75.44

(d) The number of probe vectors with Swin-Tiny.

Number	mini-1s (%)
1	73.89
4	74.01
8	74.17

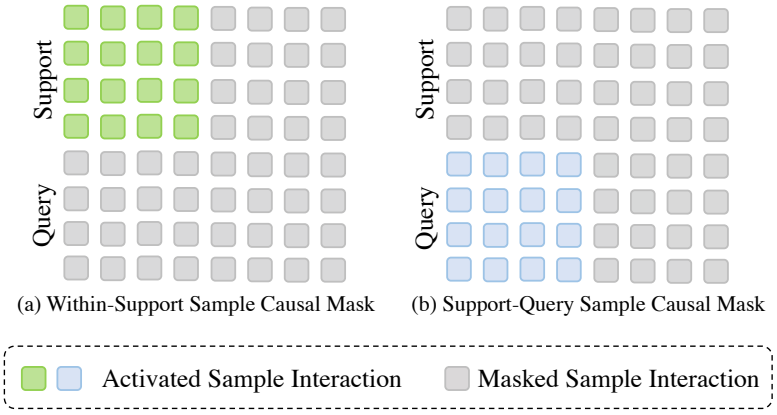
(e) Different consolidation strategies with ViT-Samll.

Method	mini-1s (%)
with averaging	75.77
without averaging	75.57
max pooling	75.63

(f) The size of knowledge pool with ViT-Samll.

Size	mini-1s (%)
10	75.50
50	75.78
100	75.74

multi-scale semantic information accounts for an improvement of 2.3%. Figure 4 shows the two alternative sample causal masks for our MetaFormer-I, where we encode the sample relationship separately. As shown in Table 16b, the ablated masks of within-support and support-query manifest sub-optimal performance, further validating that MSA with the inductive mask works not because of the introduction of extra parameters. This also underscores the complementary benefits of employing both interactions in our design on enhanced task-specific representations. Additionally, we observe that relying solely on global image features incurs a significant loss of critical information necessary for capturing discriminative relationships among samples. The results in Table 16c and Table 16d demonstrate that our model is not very sensitive to the number of task probe vectors. Also, as shown in Table 16e and Table 16f, the performance difference between various consolidation strategies is relatively marginal due to the nature of the cosine similarity-based score function employed during knowledge retrieval, and a sufficiently diverse but compact knowledge pool leads to improvements in performance. An excessively large pool risks performance degradation due to less effective consolidation.



(a) Within-Support Sample Causal Mask (b) Support-Query Sample Causal Mask

Figure 4. The alternative sample casual masks for MetaFormer-I.

G. Comparison with Other Inter-task Attention Methods

In Table 17, we compare our Patch-grained Task Attention (PTA) with other inter-task attention designs. Compared with IT-att presented in Wang et al. (2022a), we note that problem settings and underlying motivations of these modules are distinct. PTA is rooted in the domain of few-shot learning, where the paramount concern is facilitating knowledge transfer between tasks. Conversely, IT-att (Wang et al., 2022a) is grounded in continual learning, where the primary focus lies in mitigating catastrophic forgetting. While both PTA and IT-att seemingly adopt a learnable embedding for each task, their approaches to knowledge representation and utilization diverge significantly. PTA maintains a knowledge pool and captures task-specific information at the patch level. On the other hand, IT-att keeps a record of a single key and a single bias to store all past knowledge, relying on importance-based regularization to enforce task proximity and combat forgetting. Our analysis reveals that the use of a global task vector leads to a performance decrease of 0.97%. PTA demonstrates a 1.3% improvement over IT-att, which can be attributed to greater flexibility and expressiveness, aligning more closely with the objective of knowledge transfer in few-shot learning.

Table 17. Comparison results of different inter-task attention architecture design strategies. We report 5-shot accuracy on *mini*ImageNet for different choices.

Method	BackBone	Acc. (%)
IT-att (Wang et al., 2022a)	*ViT-Small*	$88.70_{\pm 0.50}$
Global task vector	*ViT-Small*	$89.05_{\pm 0.48}$
PTA	*ViT-Small*	$\mathbf{90.02}_{\pm \mathbf{0.44}}$

H. Computational Analysis

H.1. Inference Time Comparison

We have conducted a detailed comparative analysis of the computational efficiency between our MetaFormer and other state-of-the-art inductive methods for 5-way 1-shot and 5-way 5-shot scenarios on the miniImageNet, as presented in Table 3. The evaluation of inference latency is conducted on an NVIDIA RTX A6000 GPU. For FewTURE, we set the optimal inner loop steps to 15 for 5-way 1-shot and to 20 for 5-way 5-shot. Our analysis reveals that MetaFomer achieves superior computational efficiency compared to the methods evaluated. Note that previous works (Bendou et al., 2022; Qi et al., 2021; Zhu & Koniusz, 2023) utilize an inference-time augmentation technique, which involves performing inference 30 times for each randomly cropped augmented sample and then averaging the features for the final prediction. To ensure a fair comparison of inference-time latency with these state-of-the-art methods, we pre-extract features for both protoLP (Zhu & Koniusz, 2023) and MetaFormer and calculate the inference time in the 5-way 5-shot scenario. The results exhibit the superior computational efficiency of MetaFormer-A, with an inference time of 34.44 ms compared to protoLP's 40.61 ms, which is hindered by its reliance on time-consuming label propagation.

H.2. Training time Comparison

We further compare the training clock times between our MetaFormer-I and FewTURE on *mini*ImageNet in the 5-way 1-shot configuration. The results, presented in Table 18, indicate that our approach incurs a slightly higher training time. Since MetaFormer-I enhances adaptability to diverse tasks without the need for compute-intensive inference-time fine-tuning, this capability potentially leads to long-term savings in computational resources.

Table 18. Training time comparison on *mini*ImageNet.

Method	Training Clock Time
FewTURE	6.0 hours (200 epochs)
MetaFormer-I	8.7 hours (200 epochs)

H.3. The Number of Parameters Comparison

In Table 19, we present a selection of representative and state-of-the-art methods, detailing their number of backbone parameters and the total number of parameters. Simply increasing more parameters by changing the backbone architecture

does not necessarily lead to better performance. We observe that FewTURE (Hiller et al., 2022) and HCTransformers (He et al., 2022b), despite possessing a larger number of parameters, markedly lag behind the proposed MetaFormer. We also conduct a comparative analysis with an ablated version, achieved by naively augmenting the number of layers in ViT-Small to make it comparable with the proposed MetaFormer. The results presented substantiate that merely increasing parameters cannot fully address the challenges inherent in few-shot learning. In fact, such augmentation may even elevate the risk of overfitting. It's crucial to demonstrate that our enhancements are not merely due to an increased parameter count. MetaFormer is cost-effective, achieving remarkable performance gains over the previous meta-learning SOTA FewTURE (Hiller et al., 2022) with only a modest increment in parameter count. Also, note that the conversion from inductive to autoregressive version leads to no extra parameters, further emphasizing its efficiency.

Table 19. Comparsion of state-of-the-art methods with the number of parameters.

Method	Backbone	≈ # Params	# Total Params	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot
FEAT (Ye et al., 2020)	*ResNet-12*	12.4 M	14.1 M	66.78±0.20	82.05±0.14
FEAT (Ye et al., 2020)	*WRN-28-10*	36.5 M	38.1 M	65.10±0.20	81.11±0.14
FewTURE (Hiller et al., 2022)	*ViT-Small*	21 M	22 M	68.02±0.88	84.51±0.53
FewTURE (Hiller et al., 2022)	*Swin-Tiny*	28 M	29 M	72.40±0.78	86.38±0.49
HCTransformers (He et al., 2022b)	3×*ViT-Small*	63 M	63 M	74.74±0.17	89.19±0.13
SMKD (Lin et al., 2023)	*ViT-Small*	21 M	21 M	74.28±0.18	88.89±0.09
ViT with more layers	*ViT-Small*	21 M	25.2 M	69.75±0.71	84.12±0.56
MetaFormer-I (Ours)	*ViT-Small*	21 M	24.5 M	**75.78±0.71**	**90.02±0.44**
MetaFormer-A (Ours)	*ViT-Small*	21 M	24.5 M	**84.78±0.79**	**91.39±0.42**

I. Comparison with CNN-based Meta-Learning Methods

In this section, we give more in-depth discussions concerning prior research in the realm of meta-learning that incorporates vision transformers as their foundation architectures. We posit that the challenge of architectural inconsistency partially accounts for the limited research in the realm of meta-learning grounded on ViT. In the Table 20, we adapt FiLM, a technique commonly employed in CNN-based meta-learning for task adaptation through conditioned batch normalization (Requeima et al., 2019; Oreshkin et al., 2018), into layer normalization layers of ViT for task conditioning. As shown in the table, our experiments reveal a performance drop when ViT was applied with FiLM. Another key challenge is the increased parameter requirement of ViT. FewTURE (Hiller et al., 2022), as expounded in the Related Work section, is the pioneering work that tailors to ViT via inner-loop token importance reweighting, and addresses the second challenge via self-supervised pre-training on the meta-training dataset. Our approach, empowering sample-to-sample and task-to-task interaction, further improves the accuracy substantially.

Table 20. Comparison results with different meta-learning approaches for Vision Transformer on the *mini*Imagenet.

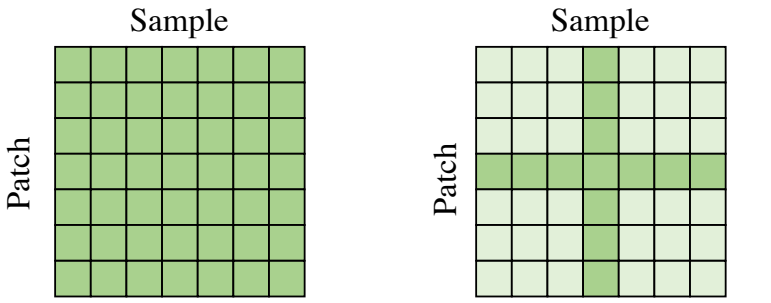
Method	BackBone	5-way 1-shot
Vanilla ViT	*ViT-Small*	69.03±0.71
ViT+FiLM	*ViT-Small*	58.75±0.73
MetaFormer-I	*ViT-Small*	**75.78±0.71**
MetaFormer-A	*ViT-Small*	**84.78±0.79**

J. Illustration of Decoupled Patch-Sample Attention

Figure 5 compares the complexity between the joint patch-sample attention and our decoupled patch-sample attention.

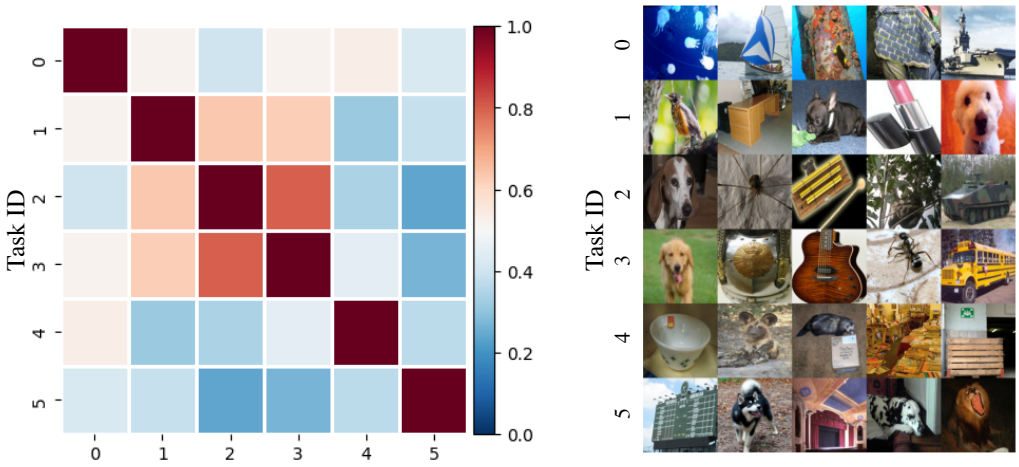
K. Additional Analysis of Task Probe Vector

As shown in Figure 6, we analyze the task probe vectors on *mini*ImageNet across different tasks sampled from meta-train and meta-test sets. The visualization effectively underscores the efficacy of the learned task probe vectors in capturing task relationships. For example, we observe a higher similarity in task features among tasks involving car tires, dogs, and long-legged animals. This demonstrates MetaFormer's capability in discerning and utilizing task dynamics.



(a) Joint Patch-Sample Attention $O(((NK+M)L)^2)$ (b) Decoupled Patch-Sample attention $O((NK+M)\cdot L^2+(NK+M)^2\cdot \mathrm{L})$

Figure 5. **Complexity comparison with joint patch-sample attention approaches.** For a N-way K-shot task with M queries, our method decouples the patch-sample attention by first performing self-attention between L patches within each image to aggregate spatial information and then computing sample interactions across all patches at the same patch location to capture the similarities and variances among samples.



(a) Task probe vector similarity (b) Task visualization

Figure 6. Interpretation of task probe vector. Each task is randomly selected from *mini*ImageNet. (a) We show the similarity heatmap between task probe vectors, where deeper color means higher similarity. (b) We show the visualization of the corresponding tasks.

L. More Qualitative Results

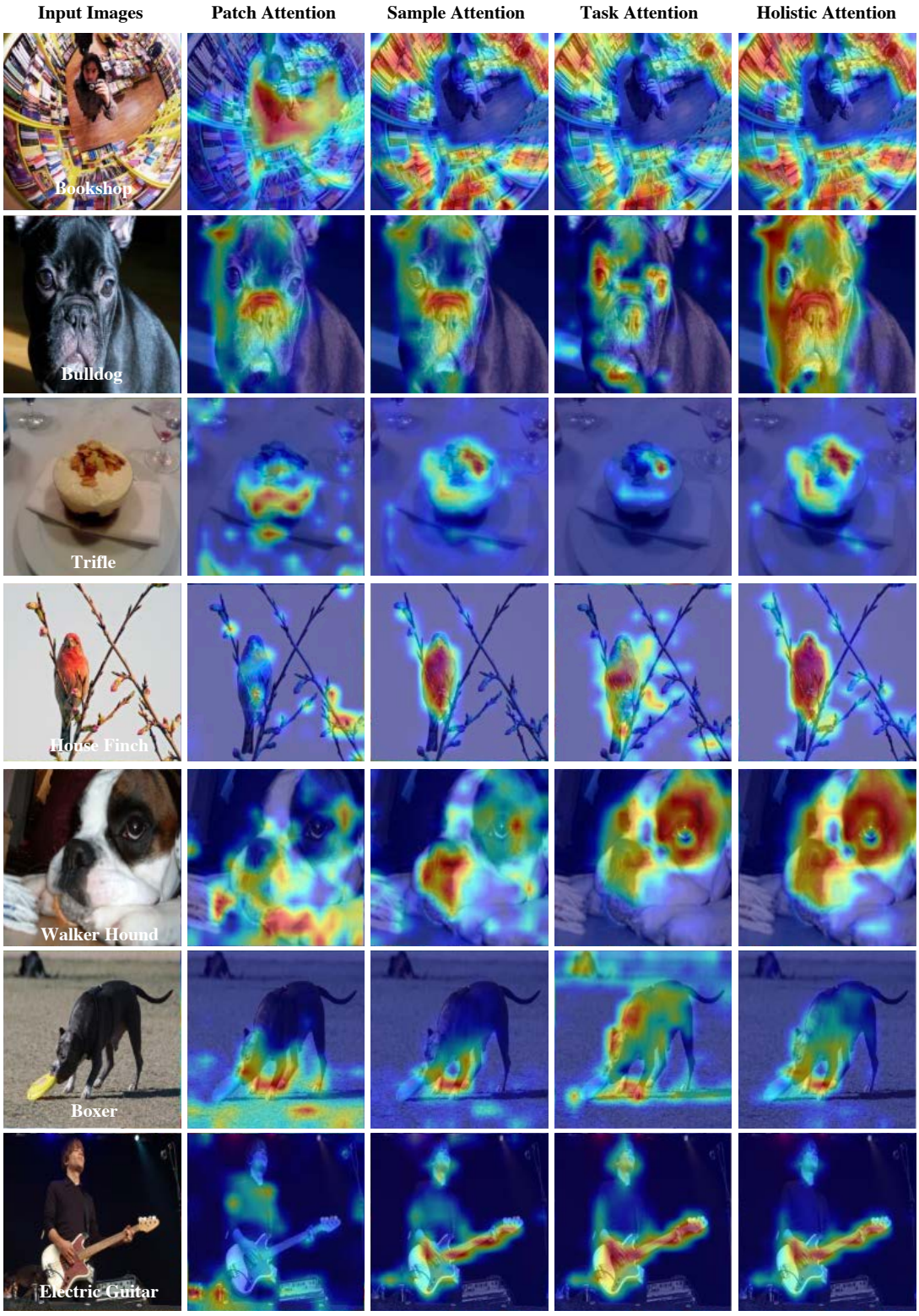


Figure 7. More visualization of our holistic attention mechanism. Three attention modules collaboratively focusing on task-specific foreground regions.