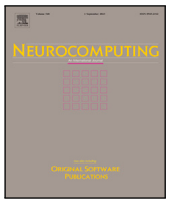
Neurocomputing 633 (2025) 129688

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



FMVP: Fine-grained Meta Visual Prompt enabled domain-specific few-shot classification

Minghui Li, Hongxun Yao *

Faculty of Computing, Harbin Institute of Technology, Harbin, 150001, China

ARTICLE INFO

Communicated by S.-J. Wang

Keywords:
Few-shot learning
Fine-grained classification
Visual prompt tuning
Cross alignment

ABSTRACT

Few-shot learning is a challenging and promising fundamental research. Inspired by recent advances in large language models (LLMs), visual prompt tuning has achieved notable performance gains in few-shot tasks by introducing only limited trainable parameters in the input space. Though effective, prompt tuning in few-shot settings heavily relies on well-initialized soft prompts and often lacks generalizability. Additionally, in certain specific fields, particularly in agriculture, there is a lack of high-precision fine-grained few-shot classification models. To our knowledge, this study is the first to employ prompt tuning for fine-grained few-shot plant disease classification (***specific to disease severity***). Specifically, we propose a novel Fine-grained Meta Visual Prompt tuning (FMVP) framework to systematically explore how visual prompts can enhance the generalizability of fine-grained few-shot domain-specific models. Firstly, a Sparsity-aware Meta Visual Prompt tuning (SMVP) sub-module is proposed to learn a universal visual prompt initialization. SMVP utilizes pixel-level optimizable visual prompts for input transformation, jointly with a novel sparsity-aware meta-learning paradigm for parameter updating, boosting generalizability to unseen classes. Secondly, a Fine-grained Cross-Alignment (FCA) module is introduced to explore intra- and inter-image relational patterns, enhancing fine-grained recognition by extracting object-level cross-image semantic discriminative features. Extensive experiments on datasets such as *mini*-ImageNet, CUB, and FPV have shown that our model outperforms state-of-the-art (SOTA) models. Our work constitutes a valuable addition to domain-specific models for practical applications.

1. Introduction

Deep learning's success has brought about significant advancements in computer vision [1–5]. However, most methods can only operate within fully supervised settings with ample available data. Specifically, supervised learning struggles to rapidly identify novel classes using just one or a few labeled samples, limiting its applicability in data-scarce domains. This has sparked growing interest in learning from a limited number of samples, a topic commonly referred to as "*few-shot learning (FSL)*" [6–8].

In reality, it is difficult to acquire samples in many real-world scenarios. For instance, plant disease recognition, a major concern for agricultural experts, represents a typical fine-grained few-shot learning problem. Rapidly detecting plant diseases is crucial for food safety and sustainability. Yet, there are few disease labels due to low incidence, high collection costs, and the requirement for manpower. Hence, a domain-specific fine-grained few-shot learning (FG-FSL) model holds significant importance and value.

The specific domain data presents several unique challenges, including high image collection costs, subtle class differences, noise sensitivity, reliance on expert annotations, low disease prevalence, and sample imbalance. These factors make fine-grained few-shot disease recognition more complex than traditional classification tasks: *limited and imbalanced data*; *bias from few-shot samples*; *few-shot prototype construction*; *fine-grained disease recognition*; *shifts in granularity*.

To tackle FG-FSL issues, we focus on high-precision few-shot and fine-grained classification methods. In the traditional domain of few-shot classification, most SOTA methods [9–12] fall into the category of metric learning. Metric-based methods learn a feature extractor, which transforms training and query samples into embeddings, then assign query embeddings to the closest class [13,14]. Building upon these baselines, many studies conduct further theoretical analysis or propose various similarity metric strategies [9–11,14–17]. In the past two years, the prompt tuning paradigm has shown greater influence in FSL. Vision-language pre-training models [18,19] pre-train on a large number of visual-textual pairs, covering almost infinite concepts in the real world,

* Corresponding author.
E-mail address: h.yao@hit.edu.cn (H. Yao).

https://doi.org/10.1016/j.neucom.2025.129688
Received 12 September 2024; Received in revised form 8 January 2025; Accepted 8 February 2025
Available online 18 February 2025

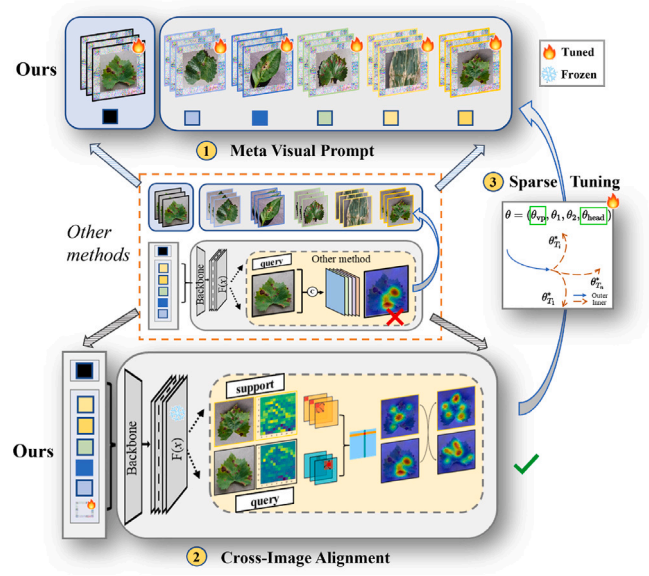


Fig. 1. A brief comparison schematic of our method with others. We briefly categorize these differences as follows: (1) Input space transformation: visual-interpretable tailored visual prompt, (2) Feature space transformation: cross-image alignment, (3) Tuning paradigm transformation: novel sparsity-aware tuning strategy. The ① MVP introduces prompt tuning into FSL to achieve good generalization. The ② cross-alignment operation performs self-attention and cross-alignment on images to focus on semantic content within and across images, ultimately obtaining more accurate cross-image discriminative features. The ③ sparsity-aware tuning algorithm updates gradients for visual prompt tuning, generating few-shot-adapted visual prompt initialization.

demonstrating impressive generalization ability on various downstream tasks [20–23]. Inspired by continuous prompt learning methods such as Prefix-Tuning [24] in natural language processing (NLP), visual prompt methods [20,25,26] achieve model fine-tuning by concatenating additional optimizable vector sequences as prompts on original input images. Though effective, prompt tuning performance is sensitive to initialization in few-shot scenarios, requiring a time-consuming process to find a good initialization, which limits the rapid adaptation ability of pre-trained models. Additionally, prompt tuning may compromise the pre-trained model's generalizability, because learnable prompts are prone to overfitting to limited training samples.

Another notable point is that existing methods typically innovate on coarse-grained general datasets, overlooking the importance of mining fine-grained features from a few classes. Mainstream works [27–33] attempt to extend meta-learning research from general to fine-grained classification by capturing discriminative parts of the entire image. SPN [34] attempts to globally align images or features through parameterized transformations. Recently, in the study [35], a relation matrix is adopted to highlight semantically relevant local features. Other works [36,37] utilize object detection methods in the image or feature space to separate the foreground and background, thus excluding background interference. While these methods are promising, little attention has been paid to the discriminability of the extracted features. Features are often extracted independently from labeled support and unlabeled query samples, leading to insufficient discriminability in few-shot scenarios. It is crucial, we believe, to align the discriminative semantic features between query and support images for computing the semantic similarity between each support-query pair.

Existing studies reveal three main limitations of current visual prompts in few-shot settings: (1) **Few-shot-adapted initialization issue**: Obtaining pixel-level visual prompt initialization for few-shot tasks is challenging, and no work has introduced visual-interpretable tailored prompts into FSL; (2) **Prompt overfitting issue**: due to the limited training samples, prompts tend to overfit and degrade generalizability, highlighting the lack of a suitable prompt tuning paradigm in few-shot scenarios; (3) **Discriminative feature mining issue**: existing prompt-based methods fail to extract fine-grained cross-image semantic discriminative features. In this paper, focusing on fine-grained disease recognition (*specific to severity*), we explore a novel **F**ine-grained **M**eta **V**isual **P**rompt tuning (FMVP) framework. It comprises two core modules: the **S**parse-Aware **M**eta **V**isual **P**rompt tuning (SMVP) module and the **F**ine-grained **C**ross-**A**lignment (FCA) module. First, we propose the SMVP module to obtain a well-initialized visual prompt and achieve strong generalizability in few-shot scenarios. SMVP applies padding-type pixel-level optimizable visual prompts to the original image for input transformation and introduces a novel two-stage meta-learning paradigm called Sparsity-Aware MAML (SA-MAML) for gradient updating. Secondly, to explore fine-grained discriminative features in few-shot scenarios, we propose the FCA module. Fine-grained discriminative feature localization across images is achieved by exploring intra- and inter-image relational patterns through tandem access to the self-attention module and cross-alignment module. A brief comparison of our proposed two modules with other methods is shown in Fig. 1. Finally, to address domain-specific FG-FSL challenges, we validate our approach in real-world tasks using a fine-grained plant disease dataset. This approach helps tackle new or rare plant disease classification problems.

Our contributions can be summarized as follows:

- We propose a Sparsity-Aware Meta Visual Prompt tuning (SMVP) module that integrates optimizable padding-type visual prompts into a novel two-stage sparsity-aware meta-learning paradigm (SA-MAML) for sparse tuning, explicitly optimizing the ability to adapt to few-shot tasks and generalize to unknown classes.
- A Fine-grained Cross-Alignment (FCA) module is introduced to extract fine-grained discriminative information across images. This module tandemly accesses the self-attention sub-module and the cross-alignment sub-module to emphasize discriminative features and align object-level cross-image semantic distributions, respectively.
- A comprehensive Fine-grained Meta Visual Prompt tuning (FMVP) joint framework is proposed to effectively address poor generalizability and discriminative representation bias issues in FG-FSL. On the classical few-shot dataset *mini*-ImageNet, CUB, and the domain-specific fine-grained dataset FPV, our approach achieves optimal performance. Our work is a valuable addition to domain-specific models for real-world applications.

The remainder of this paper is structured as follows. Section 2 summarizes related work on plant disease classification, few-shot learning, and fine-grained image recognition. Section 3 provides a detailed introduction to the proposed SMVP and FCA methods. Section 4 describes the experimental settings, ablation studies, and results analysis on different datasets. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related work

This section provides a brief overview of the related research to define and describe our proposed method. Initially, the current status of *plant disease classification* is presented (see Section 2.1), followed by an introduction to two closely related fields: *few-shot learning* (see Section 2.2) and *fine-grained recognition* (see Section 2.3).

2.1. Plant disease classification

The rapid development of deep learning has permeated various domains in agricultural scenarios. We focus on the automatic identification of plant diseases. To our knowledge, a considerable amount of research has been inclined towards developing simple deep learning networks for the automatic classification of specific plants [38–40].

Haridasan et al. [38] use an ensemble of SVM and CNNs to classify specific varieties of rice plant diseases, achieving an accuracy of 91.45% on the rice disease dataset. Using Faster R-CNN, Selvaraj et al. [39] achieve an accuracy of 94.1% in detecting banana fruit diseases. Bulakhim et al. [40] train and test AlexNet [41] and GoogleNet [42] models on 14 828 images of tomato disease leaves, achieving accuracies of 98.66% and 99.18%, respectively. Other studies tend to focus on lightweight models for classifying all detectable plant diseases [43–45]. In research [43], a lightweight and cost-effective deep learning architecture DenseNet-121 is proposed, achieving fast and efficient recognition on the PlantDoc dataset with an overall classification accuracy of only 92.5%. In a recent study [44], Abdal et al. propose a transfer learning method based on MobileNetV2 for resource-constrained environments, achieving an accuracy of up to 98.56% on their dataset.

The quantity of plant disease images used in the aforementioned research is sufficient for training deep-learning models. However, in the real world, there is often a scarcity of training data for some plant diseases due to low incidence rates and the high cost of collecting images, with only a few or a few dozen examples available, making it challenging to directly apply the above method.

2.2. Few-shot learning

Few-shot learning explores how to improve model performance with limited labeled data. Given the relevance of the algorithms to our research, we focus on two of them in particular.

Metric-based few-shot learning. In few-shot learning, most state-of-the-art (SOTA) methods fall into the category of metric learning. The metric learning-based method measures the similarity between query and support images through learned embedding functions. During testing, it classifies novel classes in the embedding space using similarity measures, where samples of the same class are closer in distance than those of different classes. Many studies are founded on these baselines, conducting further theoretical analyses or proposing various similarity measurement strategies, such as semantic orthogonality [11], Earth Mover's distance [46], Brownian distance covariance [9], negative-margin loss [16], contrastive learning embedding [47], and query-centric distance modulator [12]. In recent research, APPN [48] introduces an adaptive similarity metric, achieving optimal performance on multiple datasets. It further validates that the performance of model metrics can determine the upper limit of accuracy in few-shot classification. Abdal et al. [49] propose task attribute distance (TAD) as a novel metric to quantify task relevance, establishing a theoretical connection between task relevance and task adaptation difficulty.

Prompt tuning for few-shot tasks. In recent years, with the successive introduction of large-scale pre-trained language models such as GPT [50] and BERT [51], the prompt tuning paradigm has greatly propelled the NLP development. There have been several preliminary attempts to use visual prompts in the past two years. The earliest visual prompt method [52] is inspired by continuous prompt learning methods like Prefix-Tuning [24], extending the design concept of prompt tuning to computer vision, which directly incorporates prompt parameters into the input image, creating a prompt image called visual prompts. Based on this, Jia et al. [20] propose a method for tuning visual prompts. This method modifies a pre-trained visual Transformer by integrating selected updatable vectors into the input space. Typically based on Transformer architecture, such methods [20,25,53–55] involve appending additional optimizable vector sequences as prompts to either the original input sequence or the feature sequence at each layer of Transformer. During fine-tuning for downstream tasks, the backbone is frozen, and only the prompt vectors and the parameters of newly added modules adapted for the downstream task are optimized to achieve tuning. Differing from these methods, adding pixel-level optimizable perturbations for prompt tuning can be applied to various types of visual pre-training models (CNNs, Transformer, etc.). These methods [26,56,57] add optimizable random perturbation blocks or rectangles directly to the pixel space of the original image, independent of the model's structure. Interestingly, this approach can achieve performance comparable to fully fine-tuning models, while being more parameter-efficient and cost-effective in terms of storage. In few-shot tasks, prompt tuning demonstrates powerful performance that is incomparable to fully fine-tuning.

2.3. Fine-grained image recognition

It is required for fine-grained image classification to capture the most discriminative features amid the variations in poses of object and background, and utilize these features for image classification. These objects typically bear a striking resemblance in appearance, making them difficult to distinguish even for humans. In fully supervised fine-grained image recognition, current methods [58–60] usually employ attention mechanisms to identify the most discriminative regions and then classify them using the extracted local features. In recent research, more models [61,62] tend to partition and encode input images using self-attention-based Transformers, extracting feature representations for studying fine-grained problems. In FG-FSL, mainstream methods [27,30,31,63] aim to differentiate novel classes with only a limited number of labeled samples from the base classes. FSSA [64] proposes a spatial attention mechanism to focus on objects and suppress background noise, enabling the network to quickly learn where to attend. FYA [65] uses the cross-entropy loss between the many-hot presentation and the attention logits to optimize the model, focusing attention on key entities during fine-tuning. Although attention mechanisms have achieved remarkable success in various visual tasks, relying solely on the self-attention of query images to obtain discriminative features is often insufficient in few-shot scenarios. It is necessary to associate discriminative features between images after obtaining self-attention features.

In summary, few-shot learning and fine-grained classification have made significant strides in various visual tasks [12,46,59,60], but there is currently no effective method for fine-grained image classification with only a few labeled samples. In contrast to these approaches, we propose a novel Fine-Grained Meta Prompt tuning (FMVP) framework, which combines prompt learning and few-shot learning methods. It systematically explores how visual prompts can enhance generalizability. Meanwhile, unlike existing self-attention models, we propose a Fine-grained Cross-Alignment (FCA) module to extract object-level discriminative features across images, thereby enhancing fine-grained recognition capabilities. We expect this model to effectively tackle the FG-FSL challenge in specific domains, particularly plant disease classification.

3. Proposed method

Formal Problem Definition. Following the standard definition in prior works [13–15,66], we organize the learning process into an episodic paradigm. This paradigm progressively gathers meta-knowledge from a repository of base tasks and rapidly adapts it to novel tasks. The episodic paradigm is constructed as shown on the left side of Fig. 2. Under this setting, our goal is to train the model using labeled base classes dataset $D_{\text{base}} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N}$, and apply the classifier to unlabeled novel classes dataset D_{novel}, where x_i is an input image, y_i represents its label, N indicates the number of base class images and $D_{\text{base}} \bigcap D_{\text{novel}} = \emptyset$. To simulate few-shot testing scenarios, we reorganize all samples in $D_{\text{base}} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N}$ into a series of C-way K-shot tasks. This C-way K-shot paradigm samples C classes, extracts K labeled images per class as training samples, and then samples M samples per class from the remaining images as query images. The labeled dataset is referred to as the support set, while the unlabeled dataset is called the query set. For each episode-task T (C-way K-shot

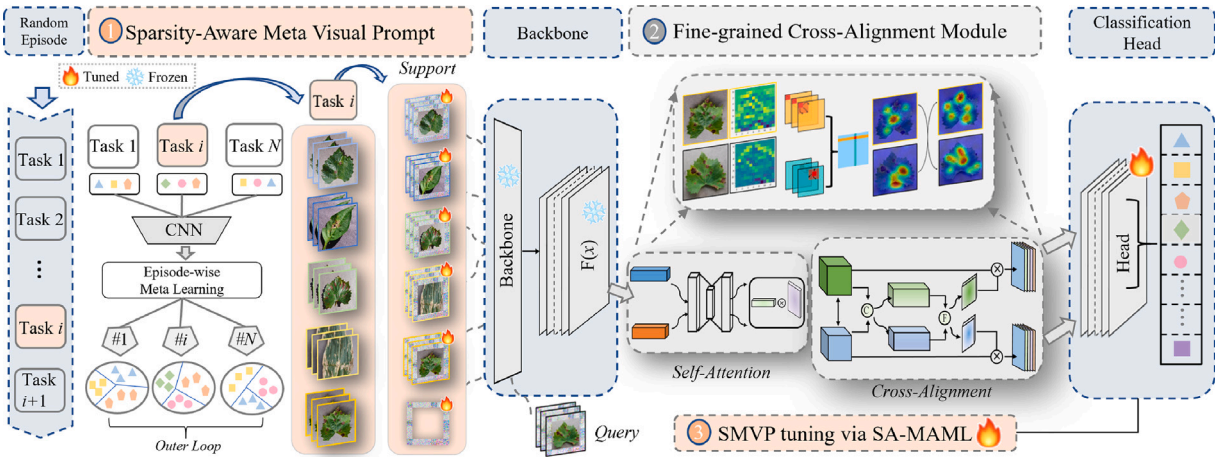


Fig. 2. Complete Fine-grained Meta Visual Prompt tuning (FMVP) framework. We carefully design episode-wise meta-learning tasks. ① SMVP introduces prompt tuning into FSL for input space transformation. Refer to Fig. 4 and Sections 3.1 for SMVP implementation details. ② FCA performs self-attention and cross-alignment operations on images to obtain more accurate cross-image discriminative features. Refer to Fig. 6 and Sections 3.3 for FCA implementation details. ③ The sparse tuning strategy SA-MAML updates the visual prompt gradients to finetune a few-shot-adapted prompt initialization. Detailed information on SA-MAML is given in Fig. 3 and Section 3.2.

M-query), the support set S and the query set Q are defined as follows:

$$S = \left\{\left(x_i^s, y_i^s\right)\right\}_{i=1}^{N_s} \left(N_s = K * C\right)$$
$$Q = \left\{\left(x_j, y_j\right)\right\}_{j=1}^{N_q} \left(N_q = 15\right), \tag{1}$$

where K refers to the number of labeled images, C denotes the number of unseen classes, and N_q is the number of query images, usually set to 15. This few-shot task is known as a C-way K-shot setting.

Overall Framework. The proposed Fine-grained Meta Visual Prompt tuning (FMVP) framework, as shown in Fig. 2, mainly consists of a Sparsity-Aware Meta Visual Prompt tuning (SMVP) module, and a Fine-grained Cross-Alignment (FCA) module. Specifically, SMVP adds pixel-level optimizable visual prompts to the input image and proposes a novel meta-gradient update strategy called Sparsity-Aware MAML (SA-MAML). The model is trained on carefully designed meta-training tasks to convert the original visual prompts into domain-generalizable directions, thereby enhancing the few-shot model's generalization ability. In FCA, we integrate the self-attention (SA) module and the cross-alignment (CA) module to explore relational patterns within and across images. This approach enables the discovery of fine-grained discriminative information by examining semantic correlations among images in few-shot scenarios. The proposed approach can be trained end-to-end, and detailed information about each component will be described in the following sections.

Sections 3.1 and 3.2 discuss the design process of visual prompt and the algorithm flow of SA-MAML (① and ③ in Fig. 2). Section 3.3 describes the implementation details of the two sub-modules in FCA (② in Fig. 2).

3.1. Sparsity-aware meta visual prompt

In Sections 3.1 and 3.2, we provide a detailed introduction to each module of SMVP, including (a) the design and selection of meta visual prompts and (b) a novel meta-learning sparsity-aware update paradigm SA-MAML.

Visual Prompt Selection. The current methods for concatenating optimizable vector sequences [20,25,53] are typically only suitable for pre-trained models with Transformer structures. In contrast, the method of adding pixel-level optimizable perturbations for prompt learning can be applied to various types of visual pre-training models. These methods [26,56,57] do not depend on the model structure but directly add optimizable random perturbation blocks or boxes to the input pixel space. During fine-tuning in downstream tasks, prompt tuning is achieved by optimizing the parameters of the perturbation. Inspired by research [52,56], we utilize visual prompts to adapt a pre-trained source model to downstream tasks without modifying any task-specific model components. Specifically, visual prompts modify input images by injecting a limited number of learnable parameters. Fig. 4 illustrates three types of pixel-level visual prompt addition methods: padding, stripe, and patch. Based on the experimental results in Section 4.3, we choose the padding-type prompt for optimal performance.

Specifically, we consider a dataset

$$\mathcal{D} = \left\{\left(x_1, y_1\right), \ldots, \left(x_n, y_n\right)\right\},$$

where x_i is the original image in $\mathcal{D}$, y_i represents its label, and n is the total number of images. The general form of input prompts is formulated as follows:

$$\tilde{x}\left(\theta_{vp}\right) = g\left(x, \theta_{vp}\right), x \in \mathcal{D} = \left\{\left(x_1, y_1\right), \ldots, \left(x_n, y_n\right)\right\}, \tag{2}$$

where $g(\cdot,\cdot)$ represents the input transformation of x with the learnable visual prompt θ_{vp}, and $\tilde{x}$ is the result after adding prompts to the original image.

We resize the input image x to a specific size $i \times i$ denoted as $s^i(x)$, where $s(\cdot)$ represents the resize operation and i is the target size. Then, we initialize the visual prompt as an 84 × 84 matrix and mask out a portion of it. Different visual prompts can be created by masking parameters of different shapes, positions, and sizes. In our selection, the prompt is a central square matrix with four adjustable outer edges. $padding$ represents the width of each outer edge. According to the results in Section 4.3, setting $padding$ to 12 yields the best performance, so we adopt it as the default value for all experiments. Finally, the input transformation operation of SMVP is described as follows:

$$\tilde{x}\left(\theta_{vp}\right) = g\left(x, \theta_{vp}\right) = s^i(x) + \theta_{vp}^{padding}, x \in \mathcal{D}. \tag{3}$$

$$\max_{\theta_{vp}} P_{\theta_{bb};\theta_{vp}}\left(y \mid x + \theta_{vp}\right). \tag{4}$$

The visual prompt parameter tuning logic. Given a frozen pre-trained model and a downstream task dataset $\mathcal{D} = \left\{\left(x_1, y_1\right), \ldots, \left(x_n, y_n\right)\right\}$, our goal is to learn a few-shot task-specific visual prompt parameterized by θ_{vp}. Let θ_{bb} denote the backbone parameters. During training, the model aims to maximize the likelihood of the correct label y :

It is important to note that gradient updates are only applied to the prompt parameters θ_{vp}, while the model parameters θ remain frozen. During evaluation, the optimized prompts are added to all query images.

$$X_{\text{test}} = \left\{x_1 + \theta_{vp}, \ldots, x_n + \theta_{vp}\right\}. \tag{5}$$

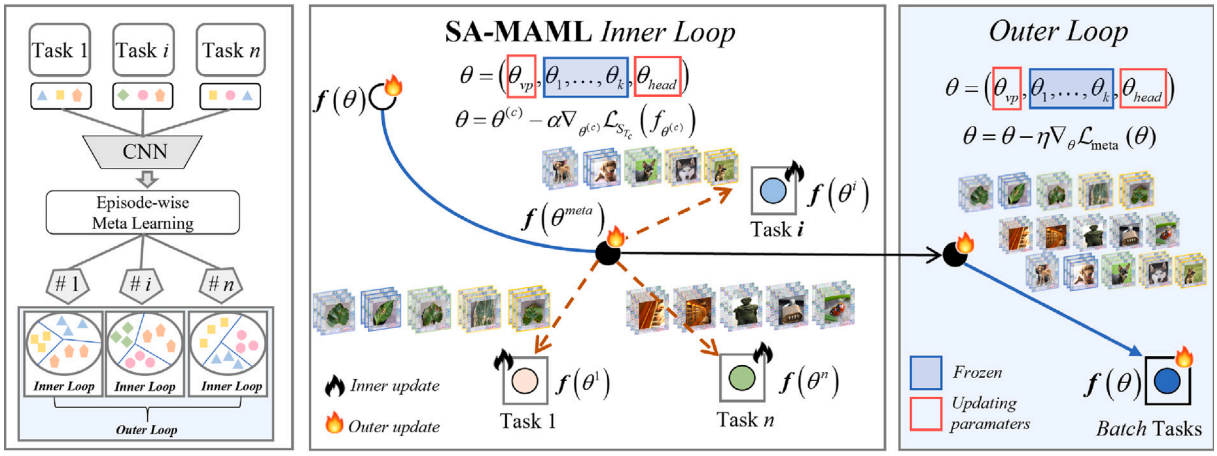


Fig. 3. Detailed schematic of the inner and outer loop updates in SA-MAML. For each meta-task, the inner loop firstly updates its task-adapted parameters. After updating all tasks in a batch, the total loss is computed, and the outer loop updates to obtain the final model parameters. Among all parameters, only θ_{vp} and θ_{head} participate in the update, while the parameters of the backbone remain fixed. The sparse updateable parameters are highlighted in red. Refer to Fig. 5 for a comparison between our method and MAML.

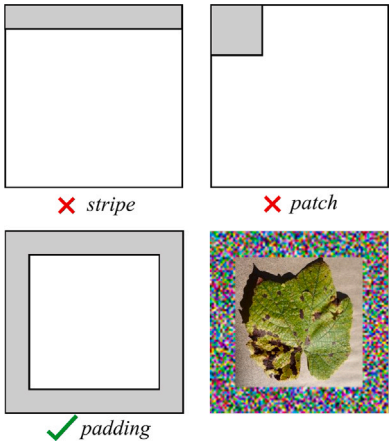


Fig. 4. Our visual prompt selection. Pixel-level visual prompts generally come in three forms: padding, stripe, and patch. We chose the padding-type visual prompt as it has been shown to deliver optimal performance.

3.2. Meta-prompt tuning via sparsity-aware MAML

After obtaining a set of meta-tasks through the episodic paradigm, we propose Sparsity-Aware MAML (SA-MAML) to learn general features among them and obtain the optimal meta-prompt initialization for few-shot generalization.

Sparsity-Aware MAML overview. Unlike other MAML-based algorithms [67,68], we propose a novel two-stage sparsity-aware meta-gradient tuning paradigm: SA-MAML, as shown in Fig. 3. It efficiently optimizes meta visual prompts to prevent prompt tuning from harming the pre-trained model's generalizability in few-shot scenarios.

Our two-stage meta-learning paradigm consists primarily of inner loop updates and outer loop updates, comprising two optimization steps. As shown in Fig. 5, the red dashed line represents the inner loop update process, while the green line represents the outer loop update process. In the inner loop, θ is initialized for each meta-training task. The updated prompt is temporarily saved and used to calculate the loss for the current task. Compared to traditional MAML [67], we add additional parameters θ_{vp} to the initialization updatable parameters set. Let θ_{head} denote the final classification layer of the network. During the inner loop update of SA-MAML, the backbone parameters are frozen, and only a limited number of parameter updates are applied to θ_{vp} and θ_{head}.

Sparsity-Aware inner loop. During SA-MAML training, we freeze the network backbone and apply sparse tuning only to the visual prompt θ_{vp} and classification head θ_{head} in the inner loop. Specifically, we first define the model as a neural network with meta-initialized parameters θ, and its model transformation process is denoted by f_θ. We randomly sample C meta-tasks $\{T_1, \ldots, T_C\}$ from the dataset D. For each task T_c, the support set S_{T_c} is used for inner loop updates, and the query set Q_{T_c} is used for outer loop updates. $\theta_i^{(c)}$ denotes the parameters of task T_c after i gradient updates. During each inner loop update, we compute:

$$\theta_n^{(c)} = \theta_{n-1}^{(c)} - \alpha \nabla_{\theta_{n-1}^{(c)}} \mathcal{L}_{S_{T_c}} \left(f_{\theta_{n-1}^{(c)}(\theta)} \right), \tag{6}$$

for n across all tasks, $\mathcal{L}_{S_{T_c}} \left(f_{\theta_{n-1}^{(c)}(\theta)} \right)$ represents the loss on the support set of T_c after $n-1$ inner loop updates. α represents the inner loop update step size.

For SA-MAML, let $\theta_{backbone} = (\theta_1, \ldots, \theta_k)$ be the k-layer main network parameters. θ_{vp} represents the visual prompt parameters, while θ_{head} represents the classification head parameters. In SA-MAML, the parameters $\theta_n^{(c)}$ of task T_c after n inner gradient updates can be represented as:

$$\theta_n^{(c)} = \begin{pmatrix} (\theta_{vp})_{n-1}^{(c)} - \alpha \nabla_{(\theta_{vp})_{n-1}^{(c)}} \mathcal{L}_{S_{T_c}} \left(f_{\theta_{n-1}^{(c)}(\theta)} \right), \\ \theta_1, \ldots, \theta_k, \\ (\theta_{head})_{n-1}^{(c)} - \alpha \nabla_{(\theta_{head})_{n-1}^{(c)}} \mathcal{L}_{S_{T_c}} \left(f_{\theta_{n-1}^{(c)}(\theta)} \right) \end{pmatrix} \tag{7}$$

In this meta-learning paradigm, only the meta-visual prompt and the network's classification layer undergo inner loop updates.

Sparsity-Aware outer loop. In the outer loop, the meta-learner aggregates knowledge from all tasks in a batch. Each meta-task calculates the loss on its query set, and the meta-learner's parameters are updated through gradient descent based on these losses.

The total loss is calculated after processing all meta-tasks in a batch:

$$\mathcal{L}_{\text{meta}}(\theta) = \sum_{c=1}^{C} \mathcal{L}_{Q_{T_c}} \left(f_{\theta_n^{(c)}(\theta)} \right), \tag{8}$$

where $\mathcal{L}_{Q_{T_c}} \left(f_{\theta_n^{(c)}(\theta)} \right)$ represents the loss on the query set after n inner loop updates.

Then, in the outer loop, θ is updated as:

$$\theta = \theta - \eta \nabla_\theta \mathcal{L}_{\text{meta}}(\theta), \tag{9}$$

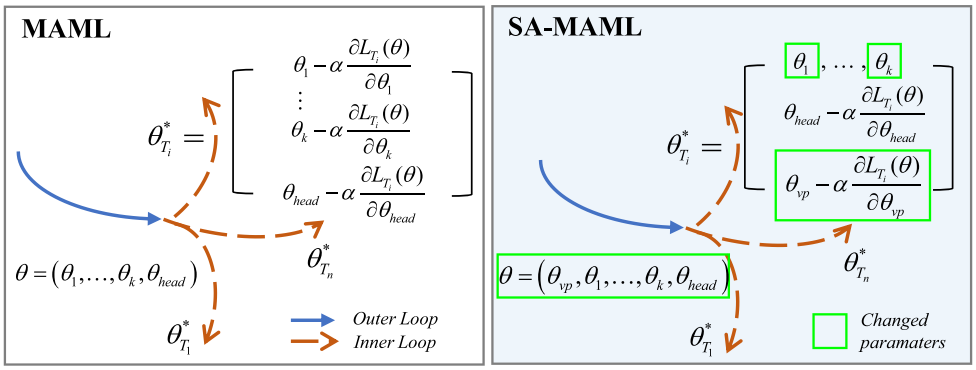


Fig. 5. Diagram of MAML and SA-MAML. Difference between MAML and SA-MAML: In MAML (left), inner loop gradient updates are applied to all parameters θ, which are then updated during the outer loop update. In SA-MAML (right), only the visual prompt parameters θ_{vp} and the classification head parameters θ_{head} are sparsely tuned in the inner loop. The green boxes highlight all the differences in gradient updates between the two algorithms.

Algorithm 1: Pseudo-code for **SA-MAML**

Require: $p(T)$: distribution over tasks
Require: support set $\mathcal{S}_{T_c}$, query set $\mathcal{Q}_{T_c}$
Require: α, η: step sizes
for $c = 1$ **to** C **do**
  Sample batch of tasks $T_c \sim p(T)$
  for *all* T_c **do**
    Sample N examples from $\mathcal{S}_{T_c}$
    Evaluate $\nabla_{\theta^{(c)}} \mathcal{L}_{\mathcal{S}_{T_c}} \left(f_{\theta^{(c)}}\right)$
    Freeze the parameters of the backbone and only update θ_{vp} and θ_{head}.
    Compute perturbation θ_{vp}:
      $\left(\theta_{vp}\right)^{(c)} - \alpha \nabla_{\left(\theta_{vp}\right)^{(c)}} \mathcal{L}_{\mathcal{S}_{T_c}} \left(f_{\theta^{(c)}}\right)$
    Compute parameters θ_{head}:
      $\left(\theta_{head}\right)^{(c)} - \alpha \nabla_{\left(\theta_{head}\right)^{(c)}} \mathcal{L}_{\mathcal{S}_{T_c}} \left(f_{\theta^{(c)}}\right)$
    Compute all tuning parameters $\theta^{(c)}$ via Eq. (7)
    Sample data from $\mathcal{S}'_{T_c}$ for meta-update
  end for
  Compute $\mathcal{L}_{\text{meta}}$ in a batch:
    $\mathcal{L}_{\text{meta}}(\theta) = \sum_{c=1}^{C} \mathcal{L}_{\mathcal{Q}_{T_c}} \left(f_{\theta_n^{(c)}(\theta)}\right)$
  Only update θ_{vp} and θ_{head}
  Update parameters $\theta^{(c)}$ via Eq. (9)
end for

where η represents the outer loop update step size.

In the outer loop, again, only the meta visual prompt and the network's classification layer are updated. Please refer to the processing flow of SA-MAML in Algorithm 1.

3.3. Fine-grained cross-alignment learning strategy

The presence of unknown labels and limited data make few-shot fine-grained classification highly challenging. Existing fine-grained methods extract features independently from labeled and unlabeled samples, resulting in limited discriminative capability. In this work, we propose a novel Fine-grained Cross-Alignment (FCA) module. It models the semantic correlation between support features and query features, thereby emphasizing fine-grained features. As shown in Fig. 6, FCA consists of two parts: the self-attention (SA) module and the cross-alignment (CA) module.

Self-Attention Module. The SA module is constructed similarly to a typical attention model, generating self-attention maps for both query and support samples. Inspired by CBAM [69], we integrate CBAM units into the feature encoder to construct a novel attention module. The simplified structure of SA is shown in the top half of Fig. 6. It focuses on discriminative information from different perspectives using channel attention and spatial attention. This module does not require separate pretraining, simplifying the training process.

Specifically, we assume the network's input feature map is $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$. We start by applying global average pooling (GAP) and global max pooling (GMP) to aggregate spatial information, resulting in two $1 \times 1 \times C$ feature maps: $\boldsymbol{I}^c_{\text{avg}}$ and $\boldsymbol{I}^c_{\text{max}}$. Next, the two feature maps are separately fed into multi-layer perceptrons (MLPs) to obtain weight coefficients between 0 and 1 using the sigmoid function σ. The weighted feature map λ^c generated by the channel attention unit can be expressed as:

$$\begin{aligned}\lambda^c &= \sigma(\text{MLP}(\text{Avg Pool}(\boldsymbol{I})) + \text{MLP}(\text{MaxPool}(\boldsymbol{I}))) \\ &= \sigma\left(\boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{I}^c_{avg}\right)\right) + \boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{I}^c_{\max}\right)\right)\right),\end{aligned} \tag{10}$$

where σ stands for the sigmoid function, $\boldsymbol{W}_1$ and $\boldsymbol{W}_0$ are the learnable parameters in the MLPs network.

The concatenated spatial attention module focuses on the important spatial features. The final weighted feature map λ^s is obtained through these steps:

$$\begin{aligned}\lambda^s &= \sigma\left(\left([\text{AvgPool}(\lambda^c \otimes \boldsymbol{I}); \text{MaxPool}(\lambda^c \otimes \boldsymbol{I})]\boldsymbol{W}_2 + b_2\right)\right) \\ &= \sigma\left(\left(\left[(\lambda^c \otimes \boldsymbol{I})^s_{\text{avg}}; (\lambda^c \otimes \boldsymbol{I})^s_{\max}\right]\boldsymbol{W}_2 + b_2\right)\right),\end{aligned} \tag{11}$$

where $\otimes$ represents element-wise multiplication, $\boldsymbol{W}_2$ and b_2 are the learnable parameters in the network.

Let SA$(\cdot)$ represent the self-attention operation. Then, the output features $\dot{\boldsymbol{h}} \in \mathbb{R}^{C\times H\times W}$ of SA is formulated as follows:

$$\dot{h} = \text{SA}(\boldsymbol{I}) = \lambda^s \otimes (\lambda^c \otimes \boldsymbol{I}). \tag{12}$$

The SA unit can be inserted into the backbone in two ways. For Conv-4 [13], the units are inserted after each convolutional layer. For ResNet-12 [13], they are inserted before the skip connections.

Cross-Alignment Module. The CA structure is shown in the lower part of Fig. 6. We define the feature map $\boldsymbol{I}_S$ extracted from the support sample and the feature map $\boldsymbol{I}_Q$ extracted from the query sample. CA generates cross-alignment maps for $\boldsymbol{I}_S$ and $\boldsymbol{I}_Q$, which are then used to weight the feature maps to obtain more discriminative feature representations. We define that after SA, the support feature map and the query feature map are denoted as $\dot{\boldsymbol{h}}_S$ and $\dot{\boldsymbol{h}}_Q$, respectively. After CA, the final support feature map and query feature map are denoted as $\dot{\boldsymbol{h}}'_S$ and $\dot{\boldsymbol{h}}'_Q$, respectively.

As shown in Fig. 6, we first design a correlation layer to compute the correlation map between $\dot{\boldsymbol{h}}_S$ and $\dot{\boldsymbol{h}}_Q$, and then use it to guide the generation of the cross-alignment map. To do this, we first reshape $\dot{\boldsymbol{h}}_S$ and $\dot{\boldsymbol{h}}_Q$ into $\mathbb{R}^{C\times m}$, where $\dot{\boldsymbol{h}}_S = \left[\vec{h}^1_S, \vec{h}^i_S, \ldots, \vec{h}^m_S\right]$ and $\dot{\boldsymbol{h}}_Q = \left[\vec{h}^1_Q, \vec{h}^i_Q, \ldots, \vec{h}^m_Q\right]$,

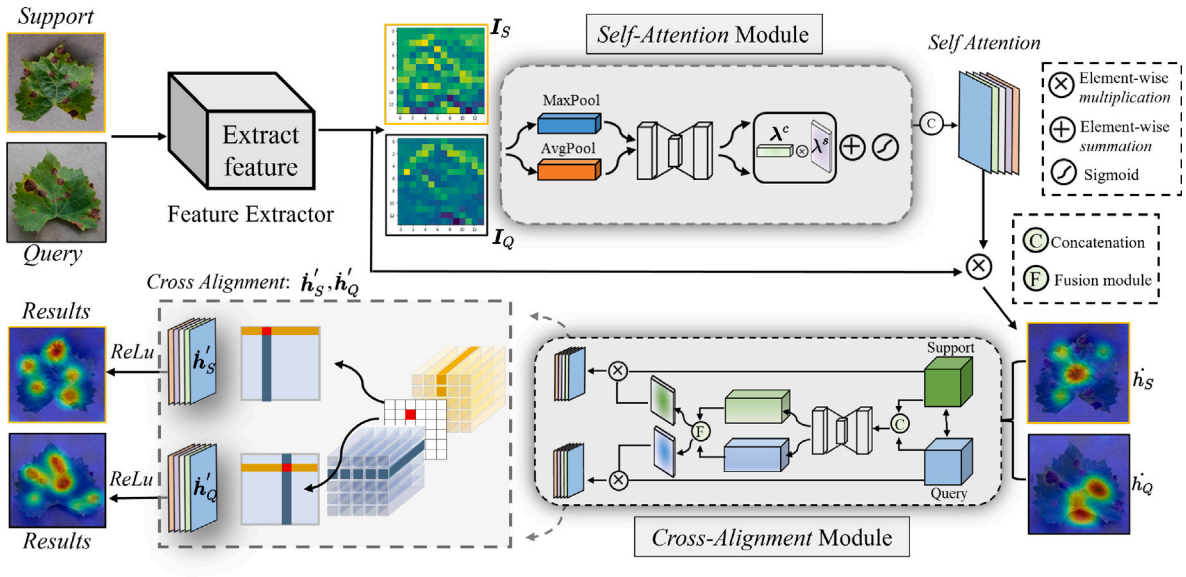


Fig. 6. The Fine-grained Cross-Alignment (FCA) module. FCA includes self-attention (SA) and cross-alignment (CA) modules. The upper part shows the self-attention module, weighting internal features with channel and spatial attention sub-modules to produce self-attention feature maps: $\dot{h}_S$ and $\dot{h}_Q$. Subsequently, these features are fed into the lower cross-alignment module for inter-image feature alignment, producing accurate cross-alignment feature maps: $\dot{h}'_S$ and $\dot{h}'_Q$.

with m ($m = H \times W$) being the number of spatial positions on the feature map. $\vec{h}^i_S$ and $\vec{h}^i_Q$ are the feature vectors of the ith spatial positions in $\dot{h}_S$ and $\dot{h}_Q$, respectively. The CA module calculates the cosine distance between the two sets of features to obtain a semantic relevance matrix:

$$\mathrm{CA}_{i,j} = \frac{\left\langle \vec{h}^i_Q, \vec{h}^j_S \right\rangle}{\left\| \vec{h}^i_Q \right\| \left\| \vec{h}^j_S \right\|}, i, j = 1, \ldots, m, \tag{13}$$

where $\langle \cdot, \cdot \rangle$ represents the dot product between two vectors, $\| \cdot \|$ indicates $\mathcal{L}_2$ normalization.

In this way, the matrix $\mathrm{CA}_{i,j}$ describes the local correlation between the support features and the query features.

Learning Meta-Fusion Embedding. We construct a meta-fusion layer to generate support and query cross-alignment maps, aligning semantically related positions in the correlation matrix $\mathrm{CA}_{i,j}$. Taking the support feature cross-alignment map as an example, we aim to recompute the spatial position weight coefficients of the support feature $\dot{h}_S$ to align with the query feature $\dot{h}_Q$. Specifically, each row of the semantic correlation matrix should be normalized to sum to 1, to be used as a weighted vector for the support feature $\dot{h}_S$. The formula for normalizing the semantic correlation matrix row by row is as follows:

$$\overline{\mathrm{CA}}_{i,j} = \frac{\exp\left(\mathrm{CA}_{i,j}\right)}{\sum_{t=1}^{HW} \exp\left(\mathrm{CA}_{t,j}\right)} \tag{14}$$

The final aligned support feature $\dot{h}'_S \in \mathbb{R}^{C \times HW}$ is obtained by calculating the matrix multiplication between $\overline{\mathrm{CA}}$ and the transpose of the support feature $\dot{h}_S$ (The calculation for the query feature $\dot{h}'_Q$ is similar):

$$\dot{h}'_S = \left(\overline{\mathrm{CA}} \cdot \dot{h}_S{}^T\right)^T. \tag{15}$$

4. Experimental result and analysis

In this paper, we concentrate on FG-FSL tasks, specifically exploring disease classification that is precise to the severity levels. In this section, we answer the following questions:

- **Q1.** Are all the modules in the ablation experiments necessary and advanced? (Section 4.3)
- **Q2.** How does our method perform on coarse-grained and fine-grained tasks in the general domain? (Section 4.4)
- **Q3.** What roles do different modules play in tasks of different granularities? Which one plays the dominant role? (Sections 4.3 and 4.4)
- **Q4.** Is the model effective for tasks with extremely fine granularity in the specific agricultural domain? (Section 4.5)

Table 1
Datasets.

Dataset	Total(#)	Base(#)	Val(#)	Novel(#)
mini-ImageNet	100	64	16	20
CUB	200	100	50	50
FPV	59	30	14	15

4.1. Experiments

Dataset Overview. To evaluate our model effectively, we conduct experiments on common few-shot datasets like *mini*-ImageNet [70] and Caltech-UCSD-Birds (CUB) [71], and also apply it to a fine-grained plant disease classification dataset, Fine-PlantVillage (FPV).

mini-ImageNet is a subset of ImageNet extracted for few-shot tasks and is a commonly used coarse-grained dataset. *mini*-ImageNet consists of 100 different classes with images sized at 84 × 84 × 3. Following the common split method in previous research [13,14,72], we divide the dataset into 64 base classes, 16 validation classes, and 20 novel classes.

CUB has smaller inter-class differences and higher recognition difficulty, making it a commonly used fine-grained few-shot dataset. CUB contains 200 different species of birds. Similarly, following the standard split method [13,16,73], we divide the dataset into 100 base classes, 50 validation classes, and 50 novel classes.

FPV is a fine-grained plant disease dataset that further refines the categories based on PlantVillage (PV). It contains 10 plants and 61 plant diseases, with labels refined to include disease severity levels in the format "*plant-disease-severity*", as shown in Fig. 11. Due to two diseases having fewer than five samples, we excluded them from the analysis. For the fine-grained classification task, we select 30 diseases as base classes, 14 as validation classes, and 15 as novel classes. The splitting

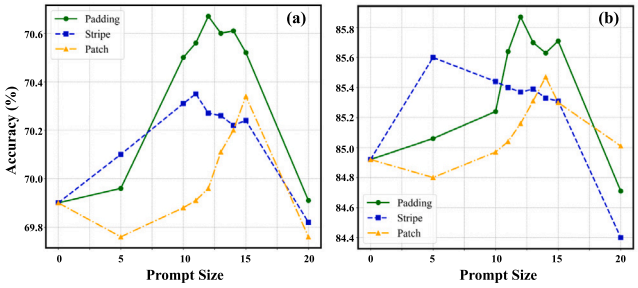


Fig. 7. Ablation study on visual prompts. The backbone ResNet-12 is pre-trained on the *mini*-ImageNet base classes and validated for accuracy on the novel classes. (a) Experimental results for 1-shot settings. (b) Experimental results for 5-shot settings. The padding-type prompt: border width = *prompt size*. The stripe-type prompt: fixed length and width = *prompt size*. The patch-type prompt: square with sides = *prompt size*.

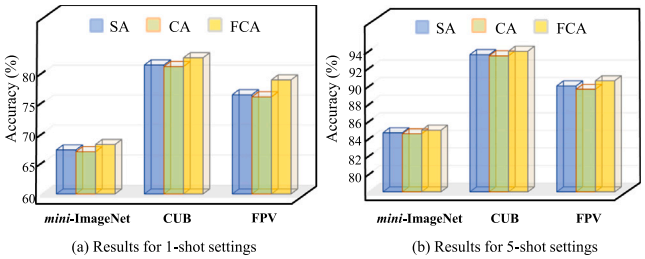


Fig. 8. Results of ablation experiments for SA, CA, and FCA. Blue, green, and yellow respectively represent experimental results using only SA, CA, and FCA. (a) Experimental results on three datasets under 1-shot settings. (b) Experimental results on three datasets under 5-shot settings.

of the three datasets is illustrated in Table 1.

Benchmarking Methods. In this paper, we compare our method with earlier classical methods and recent state-of-the-art (SoTA) methods, such as QCDM [12], GLFA [11], STANet [10], DeepEMD [46], DeepBDC [9], EMO [74], and IAM [75]. These methods have been evaluated on multiple few-shot datasets and have achieved optimal performance.

Ablation Studies. In ablation studies, we compare the performance differences of SMVP, SA-MAML, and FCA when they appear independently and in combination, validating the effectiveness of each module.

Comparison with State-of-the-art. In the method comparison, we conduct experiments on the above datasets with a unified setting of 5-way 1-shot or 5-way 5-shot. On datasets of different granularities, we conduct quantitative and qualitative analyses of the model's performance. Specifically, for domain-specific tasks, we visualize visual prompts, confusion matrices, cross-alignment heatmaps, t-SNE feature reduction graphs, etc., to assist in the analysis.

4.2. Implementation details

Training settings. We use ResNet-12 [13] as the feature extractor backbone. The ResNet-12 backbone's pretraining parameters are trained using all labeled base class data. During meta-tuning, the backbone's parameters remain frozen, and only a few parameters of the visual prompt and the classification head are sparsely fine-tuned using SA-MAML. All images are resized to 84×84, momentum is set to 0.9. The initial learning rate is 0.001, halved every 20 epochs. We train each dataset for 500 epochs and select the best-performing model. Experiments are conducted using PyTorch on an NVIDIA 2080 Ti GPU.

Evaluation metrics. The model's performance is evaluated using standard metrics [9,12,14]. We evaluate the model using 2000 randomly sampled episodes (15 query images per class) in n-way k-shot settings and report the average top-1 accuracy (%). We report the classification accuracy of different methods on the *mini*-ImageNet, CUB, and FPV datasets under 5-way 1-shot and 5-way 5-shot settings.

4.3. Ablation studies

In this section, we conduct experiments with ResNet-12 as the backbone under the 5-way 5-shot setting. As shown in Table 2, to further validate the effectiveness of SMVP, SA-MAML, and FCA, we apply these modules individually or in combination to the baseline model and conduct experiments on three datasets. For the design of visual prompts, we also conduct multiple ablation experiments for prompt types and sizes.

In Table 2, the mark ($\surd$) is placed in the corresponding cell if the module is added, otherwise, it remains blank. Marks ($\surd$) in multiple cells indicate the simultaneous use of multiple modules. SMVP* refers to an incomplete SMVP module using MAML, while SMVP refers to the complete version using the proposed SA-MAML. FCA* denotes an incomplete version using only SA, while FCA represents the complete version incorporating both SA and CA.

The influence of Prompt Design. To investigate the effects of prompt types and sizes, we conduct ablation experiments on three different types of visual prompts. The image input size is fixed at 84×84, and the prompt size varies from 0 to 20. The backbone ResNet-12 is pre-trained on *mini*-ImageNet base classes, and the test accuracy is obtained on the novel classes. All prompt updates are driven by SA-MAML, and FCA is applied to obtain fine-grained features. As depicted in Fig. 7, the padding-type prompt achieves optimal performance with a size around 12 in both 1-shot and 5-shot settings. Both too-large and too-small prompt sizes lead to a significant decrease in performance in few-shot scenarios. We chose a padding size of 12 for the padding-type prompt, which has shown the best performance, as the default value for all experiments.

The influence of SA & CA. The FCA module consists of the SA sub-module and the CA sub-module. Fig. 8 illustrates the classification results of SA, CA, and FCA separately under 1-shot and 5-shot settings. Table 2 shows the classification results for SA and FCA when they appear alone or in combination with other modules. Models using each sub-module individually outperform those without attention modules, demonstrating the effectiveness of both. SA can adaptively fuse channel and spatial information to generate self-attention maps, while CA aligns discriminative information across images for finer feature extraction. FCA combines the advantages of foreground object enhancement and cross-image semantic alignment, resulting in optimal performance.

The influence of FCA. Row 5 in Table 2 shows the classification results with only FCA added. The experiments demonstrate that FCA can effectively improve classification performance on all three datasets. On the coarse-grained dataset, FCA achieves a classification accuracy improvement of 1.12%. On the fine-grained datasets CUB and the even finer-grained FPV, the classification accuracy improves by 1.26% and 1.62%, respectively. As the granularity becomes finer, FCA provides a greater benefit to the model. This suggests that the cross-alignment mechanism is crucial for extracting fine-grained features, especially on datasets with finer granularity. In Fig. 9, we visualize attention heatmaps for some samples from *mini*-ImageNet and CUB. The third column shows the self-attention map after SA, and the fourth column shows the cross-alignment attention map after FCA.

The influence of SA-MAML. Comparing rows 2 and 3, 6 and 8, and 7 and 9 in Table 2, the SA-MAML-driven models consistently outperform those driven by MAML in each group of comparative experiments. Comparing rows 1 and 2, 4 and 6, it can be seen that the MAML models even lead to a decrease in accuracy on some datasets. This effect may be attributed to MAML's simultaneous fine-tuning of all parameters, which could impede convergence and the acquisition of effective initial visual prompts. In the SA-MAML paradigm, only a few visual prompts and

Table 2
Ablation study of 5-way 5-shot on few-shot classification. The best results are displayed in **boldface** (mean ± S.D.%). Numbers are in percentage (%). The mark (√) indicates that the module is used.

Baseline	SMVP* (original MAML)	SMVP (SA-MAML)	FCA* (only SA)	FCA (SA+CA)	*mini*-ImageNet	CUB	FPV
√					83.80 ± 0.31	92.74 ± 0.27	89.02 ± 0.38
√	√				83.62 ± 0.30	93.11 ± 0.33	89.48 ± 0.33
√		√			85.08 ± 0.44	93.53 ± 0.26	89.77 ± 0.30
√			√		84.73 ± 0.28	93.61 ± 0.30	90.06 ± 0.27
√				√	84.92 ± 0.35	94.00 ± 0.31	90.64 ± 0.40
√	√		√		84.71 ± 0.36	93.72 ± 0.29	90.13 ± 0.44
√	√			√	84.98 ± 0.30	94.04 ± 0.32	90.40 ± 0.41
√		√	√		85.39 ± 0.41	94.01 ± 0.26	90.76 ± 0.29
√		√		√	**85.87 ± 0.28**	**94.64 ± 0.28**	**91.18 ± 0.26**

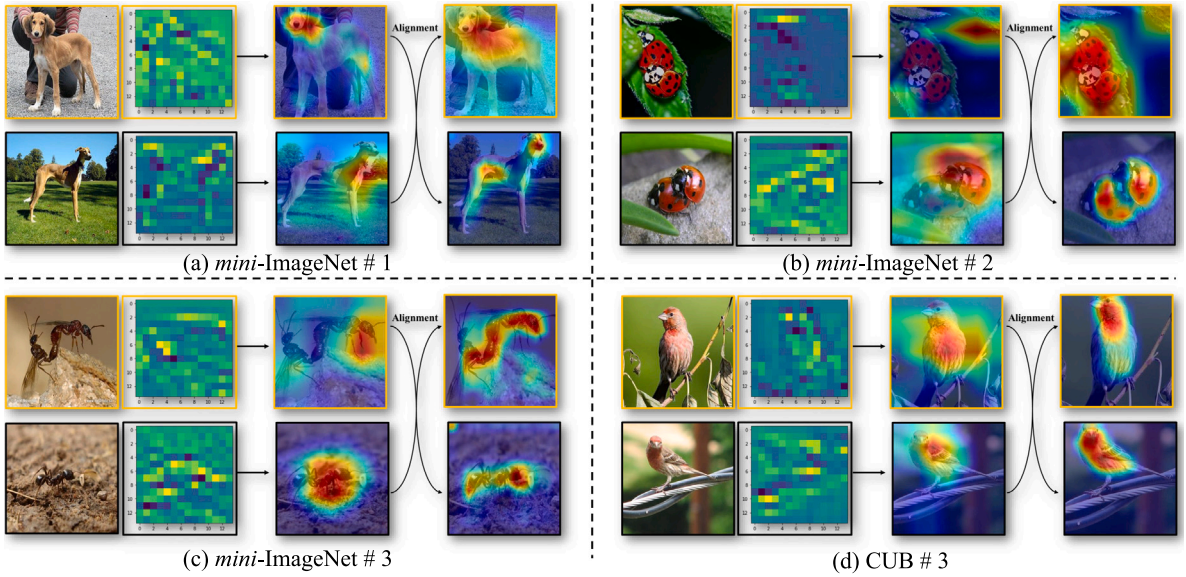


Fig. 9. Visualization of attention heatmaps before and after applying CA. (a)(b)(c) visualize attention heatmaps on the coarse-grained dataset *mini*-ImageNet. (d) visualizes attention heatmaps on the fine-grained dataset CUB. The third column displays self-attention heatmaps, while the fourth column shows cross-alignment heatmaps after applying CA.

classification head parameters are updated during inner and outer loop gradient updates. This sparse tuning paradigm improves the model's generalization ability.

The influence of SMVP. In Table 2, comparing rows 1 and 3, SMVP shows significant performance gains on all three datasets. It improves by 0.79% and 0.75% on the fine-grained datasets CUB and FPV, respectively. For the coarse-grained *mini*-ImageNet, the accuracy is greatly increased by 1.12%. This is in contrast to the gain effect brought by FCA. We infer that in coarse-grained tasks, SMVP plays a more critical role than SA-MAML. SMVP utilizes the SA-MAML paradigm to obtain well-adapted visual initializations for few-shot tasks, performing well in tasks with general granularity, while FCA's impact is relatively minor due to the scarcity of difficult-to-classify fine-grained samples. Fig. 10 visualizes the padding-type visual prompts trained under different settings on three datasets. All prompts are generated through sparse updates based on SA-MAML.

As shown in Table 2, the combination of SMVP and FCA achieves the best performance. In conclusion, a series of ablation experiments demonstrate the effectiveness and strong generalization ability of the proposed modules.

4.4. Comparison with state-of-the-art

To evaluate our method, we compare it with classic and SOTA few-shot algorithms [9–12,16,46,67,73,74,79–81] on the general few-shot datasets, *mini*-ImageNet and CUB, using n-way k-shot settings.

Backbone Selection. In few-shot learning, mainstream methods commonly use Conv-4 [13] and ResNet-12 [13] as backbone networks. We have also conducted comparative experiments using these two backbone networks, comparing classical methods such as Relation N [15], Meta-Baseline [73], Neg-Margin [16], etc., as well as SOTA methods such as STANet [10] and QCDM [12]. As shown in Table 4, our model achieves superior performance in most experiments on both backbone networks, especially showing a more significant accuracy improvement with the deeper ResNet-12. Due to our focus on solving fine-grained few-shot classification problems, which require strong feature extraction capabilities, all experiments in the following sections are conducted based on ResNet-12.

Table 3 shows the experimental results on *mini*-ImageNet and CUB under 5-way 1-shot and 5-way 5-shot settings. Metric-based methods outperform optimization-based and generation-based methods in accuracy, leading us to focus on comparing more metric-based methods.

Results for coarse-grained *mini*-ImageNet. In coarse-grained *mini*-ImageNet experiments, our method outperforms other SOTA methods [9–12,46,74] in both 1-shot and 5-shot settings. In 5-shot tasks, compared with the latest STANet [10], IAM [75], and QCDM [12], the classification accuracy improved by 0.99%, 1.01%, and 0.96%, respectively. In 1-shot tasks, our method achieves a 0.83% improvement over the second-best STANet [10]. Combining the ablation experiments in Table 2, we infer that SMVP is more suitable for general-grained tasks, while FCA's impact is more pronounced in fine-grained tasks. In coarse-grained tasks with significant inter-class differences, SMVP boosts few-shot generalization by adapting prompts to few-shot tasks, while FCA struggles due to the easy classification properties of coarse

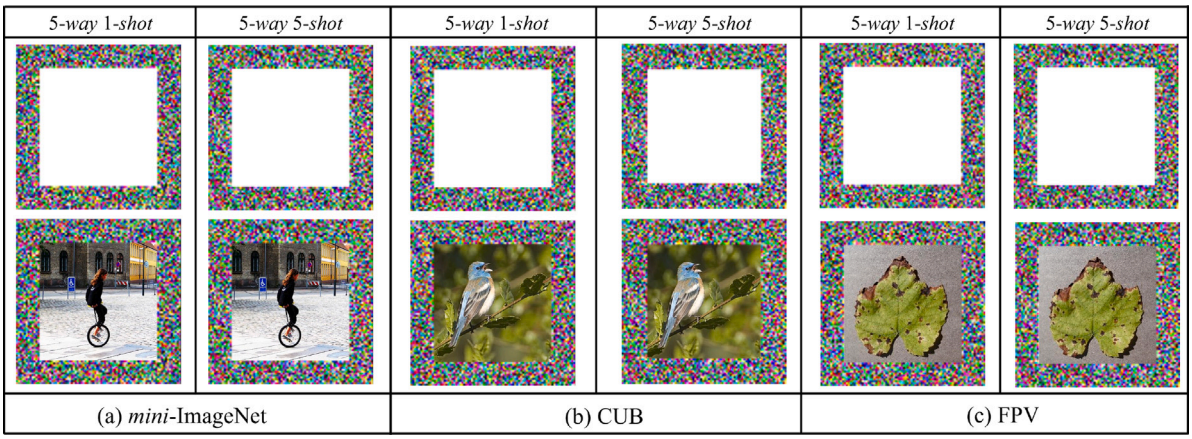


Fig. 10. Visualization of padding-type visual prompts trained under different settings on various datasets. All prompts are generated through sparse updates using SA-MAML. The top part of each dataset shows the task-specific visual prompt, while the bottom part shows the transformed input image.

Table 3
Comparison of the state-of-the-art few-shot classification algorithms on the *mini*-ImageNet and CUB dataset. Numbers are in percentage (%). The best results are highlighted in **bold** (mean ± S.D.%).

Methods	*mini*-ImageNet		CUB	
	5way1shot	5way5shot	5way1shot	5way5shot
Optimization-based				
MAML [67] (2017)	57.40 ± 0.47	72.42 ± 0.65	70.44 ± 0.55	85.50 ± 0.33
E^3BM [76] (2020)	64.45 ± 0.34	81.04 ± 0.53	78.22 ± 0.61	89.34 ± 0.35
EMO [74] (2023)	69.15 ± 0.34	84.13 ± 0.25	–	–
Generation-based				
MVT [77] (2020)	–	67.67 ± 0.70	–	80.33 ± 0.60
TriNet [78] (2019)	58.12 ± 1.37	76.92 ± 0.69	69.61 ± 0.46	84.10 ± 0.30
Metric-based				
Baseline [79] (2019)	60.00 ± 0.44	80.55 ± 0.31	71.85 ± 0.46	88.09 ± 0.25
Baseline++ [79] (2019)	63.25 ± 0.44	81.67 ± 0.30	75.25 ± 0.45	89.85 ± 0.23
Meta-Baseline [73] (2020)	64.17 ± 0.45	81.41 ± 0.31	78.16 ± 0.43	90.04 ± 0.23
Neg-Margin [16] (2020)	61.70 ± 0.46	78.03 ± 0.33	78.14 ± 0.46	90.00 ± 0.24
FEAT [80] (2020)	66.78 ± 0.20	82.05 ± 0.14	77.53 ± 0.83	89.79 ± 0.28
BML [81] (2021)	67.04 ± 0.63	83.63 ± 0.29	77.21 ± 0.63	90.45 ± 0.36
DeepEMD [82] (2020)	65.91 ± 0.82	82.41 ± 0.56	75.65 ± 0.63	88.69 ± 0.50
MCL [83] (2022)	67.36 ± 0.20	83.63 ± 0.20	–	–
DeepBDC [9] (2022)	67.83 ± 0.43	84.45 ± 0.29	79.01 ± 0.42	90.42 ± 0.17
SetFeat12 [84] (2022)	68.32 ± 0.62	82.71 ± 0.46	79.60 ± 0.80	90.48 ± 0.44
DeepEMD v2 [46] (2022)	68.77 ± 0.29	84.13 ± 0.53	–	–
IAM [75] (2023)	67.95 ± 0.19	84.86 ± 0.13	78.28 ± 0.22	90.72 ± 0.12
STANet [10] (2023)	69.84 ± 0.47	84.88 ± 0.30	80.46 ± 0.47	90.88 ± 0.30
GLFA [11] (2023)	67.25 ± 0.36	82.80 ± 0.30	76.52 ± 0.37	90.27 ± 0.38
MB+QCDM [12] (2024)	66.76 ± 0.23	84.91 ± 0.18	83.54 ± 0.29	93.01 ± 0.21
HFCR [85] (2024)	–	–	84.39 ± 0.19	93.40 ± 0.11
Ours	**70.67 ± 0.36**	**85.87 ± 0.28**	**85.50 ± 0.40**	**94.64 ± 0.28**

features. In our main comparison, methods like STANet [10] and GLFA [11] improve accuracy by proposing novel attention mechanisms, while QCDM [12], DeepEMD [46], and DeepBDC [9] focus on optimizing metric schemes. Their improvements are limited in terms of few-shot-adapted prototype initialization and model generalization. The experiments demonstrate that the carefully designed SMVP is more suitable for coarse-grained few-shot tasks than classical metric-based and optimization-based models.

Results for fine-grained CUB. Fine-grained images exhibit higher intra-class variance and lower inter-class variance, posing significant challenges in both supervised learning and few-shot learning domains. On CUB, our model achieves 85.50% accuracy under 1-shot settings, outperforming HFCR [85] by 1.11%, another model focusing on attention mechanisms. In 5-shot settings, our model improves by 1.24%, 1.63%, and 3.76% compared to HFCR [85], QCDM [12], and STANet [10], respectively. Combining the ablation experiments (Table 2) and comparison experiments (Table 3) across different granularities, we observe that FCA achieves higher performance gains on finer granularity datasets. We believe this is because in fine-grained tasks, extracting discriminative information is more critical than enhancing generalization ability. FCA explores relationships within and across images to extract fine-grained object-level semantic relationships, thereby enhancing fine-grained recognition capabilities.

Finds of general few-shot classification. Combining ablation and accuracy experiments, we can summarize the key findings of the model as follows:

- **(1)** As the granularity level becomes finer, the roles of the two sub-modules, SMVP and FCA, in FMVP also change.
- **(2)** For coarse-grained tasks, SMVP's few-shot-adapted generalizability obtained through sparse meta visual prompt initialization is more critical, while FCA plays a supporting role.
- **(3)** For fine-grained tasks, FCA obtains cross-image object-level fine-grained information through self-attention foreground enhancement and cross-semantic alignment, achieving higher gains compared to SMVP.

Table 4
Few-shot results with different settings of backbones (Conv-4 and ResNet-12). The best results are displayed in **boldface** (mean ± S.D.%). Numbers are in percentage (%).

Methods	Backbones	*mini*-ImageNet		CUB	
		5way1shot	5way5shot	5way1shot	5way5shot
Relation N [15] (2018)	Conv-4	49.69 ± 0.43	68.14 ± 0.35	–	–
	ResNet-12	54.12 ± 0.46	71.31 ± 0.37	73.22 ± 0.48	86.94 ± 0.28
Baseline [79] (2019)	Conv-4	46.06 ± 0.39	65.83 ± 0.35	47.73 ± 0.41	68.77 ± 0.38
	ResNet-12	60.00 ± 0.44	80.55 ± 0.31	71.85 ± 0.46	88.09 ± 0.25
Baseline++ [79] (2019)	Conv-4	51.16 ± 0.43	67.99 ± 0.36	62.01 ± 0.49	77.72 ± 0.36
	ResNet-12	63.25 ± 0.44	81.67 ± 0.30	75.25 ± 0.45	89.85 ± 0.23
Meta-Baseline [73] (2020)	Conv-4	51.35 ± 0.42	66.99 ± 0.37	58.98 ± 0.47	75.77 ± 0.37
	ResNet-12	64.17 ± 0.45	81.41 ± 0.31	78.16 ± 0.43	90.04 ± 0.23
Neg-Margin [16] (2020)	Conv-4	51.15 ± 0.42	67.32 ± 0.35	64.08 ± 0.48	80.69 ± 0.34
	ResNet-12	61.70 ± 0.46	78.03 ± 0.33	78.54 ± 0.46	90.19 ± 0.24
STANet [10] (2023)	Conv-4	57.32 ± 0.47	73.00 ± 0.37	65.57 ± 0.44	81.89 ± 0.35
	ResNet-12	69.84 ± 0.47	84.88 ± 0.30	80.46 ± 0.47	90.88 ± 0.30
MB+QCDM [12] (2024)	Conv-4	55.27 ± 0.45	72.41 ± 0.32	64.76 ± 0.48	82.19 ± 0.30
	ResNet-12	66.76 ± 0.23	84.91 ± 0.18	83.54 ± 0.29	93.01 ± 0.21
Ours	Conv-4	**57.46 ± 0.50**	**73.24 ± 0.33**	**65.94 ± 0.43**	**82.40 ± 0.30**
	ResNet-12	**70.67 ± 0.36**	**85.87 ± 0.28**	**85.50 ± 0.40**	**94.64 ± 0.28**

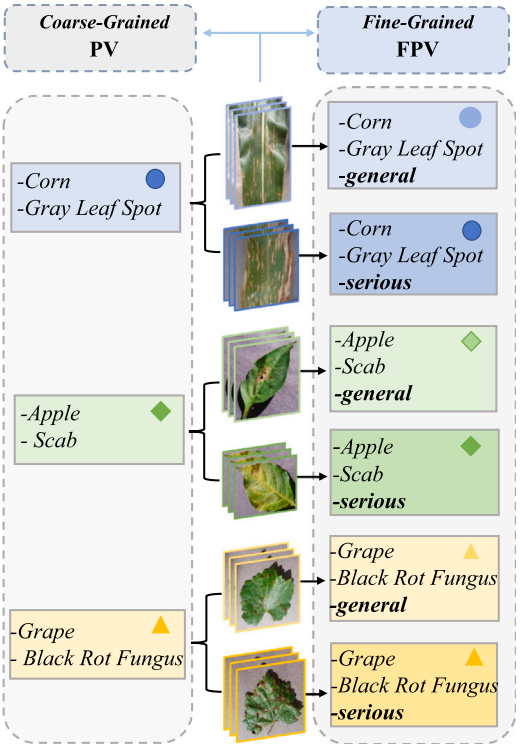


Fig. 11. Fine-grained examples from FPV. (a) The left side of the figure displays coarse-grained categories in the format "*plant-disease*", (b) while the right side displays fine-grained categories in the format "*plant-disease-severity*". Different shapes represent different diseases (e.g., ◊ = *Apple-Scab*), and colors indicate disease severity (e.g., Dark green =*Apple-Scab-serious*). For the same image, its category is different at different fine-grained levels. Our focus is on the more fine-grained disease classification on the right side.

- **(4)** Our proposed FMVP achieves optimal performance in tasks of different granularities by combining the advantages of both.

4.5. Domain-specific tasks

For specific tasks like plant disease classification, collecting a large number of samples is often impractical due to the difficulty in determining disease symptoms. This process typically requires experienced agricultural experts for labeling. Currently, there are several open-source datasets available online, but their classification granularity typically only reaches the level of disease categories. The FPV dataset, unlike existing open-source datasets that generally classify diseases at a basic level, refines disease classification to severity levels. This granularity is crucial for tackling real-world agricultural challenges.

Fig. 11 displays examples from the FPV dataset. The left side shows coarse-grained category labels in the format "*plant-disease*", while the right side shows fine-grained labels for each image in the format "*plant-disease-severity*". PV is a plant disease image dataset that contains 54,305 images of plant disease leaf samples across 14 plant species with 38 disease classes. FPV, built upon PV, further refines the categories, including 10 plants and 61 diseases, totaling 45,285 images. From Fig. 11, it can be observed that the same set of images are assigned different labels under different fine-grained requirements. Fine-grained classification in FPV poses a significant challenge, especially in classifying severity levels of diseases within the same disease category.

In this section, we will conduct a series of experiments on FPV using n-way k-shot settings. In FPV, we randomly select 30 classes as base classes and the remaining classes as validation and novel classes to perform FSL tasks.

Results for finer-grained FPV. As shown in Table 5, compared to experiments on *mini*-ImageNet and CUB, our method achieves the highest performance improvement in both 1-shot and 5-shot settings on the finer-grained FPV dataset. Specifically, in 5-shot experiments on FPV, our method outperforms QCDM [12] and DeepBDC [9] by 1.12% and 1.73%, respectively. In 1-shot experiments, our method improves the accuracy from the second-best 80.06% to 81.86%. Combining Tables 2 and 5, we can analyze the performance improvement trends across the three datasets. In 5-shot (1-shot) settings, the performance of the models on *mini*-ImageNet, CUB, and FPV datasets improves by 0.96% (0.83%), 1.03% (0.96%), and 1.12% (1.80%), respectively. We assume that the three datasets represent three different levels of granularity. It can be observed that as the granularity becomes finer, our method's advantage over other SOTA methods becomes more pronounced. In contrast, other metric-based methods, such as QCDM [12] and DeepBDC [9], do not exhibit this characteristic. Overall, our model demonstrates the most competitive performance in facing extremely fine-grained tasks in specific domains.

Padding-Type Visual Prompt Visualization. Fig. 10 visualizes the padding-type visual prompts obtained from training under different tasks. A detailed comparison reveals that our model generates distinct specialized task-adapted visual prompts for different datasets and different k-shot settings. In Fig. 10(c), the visualization results for FPV

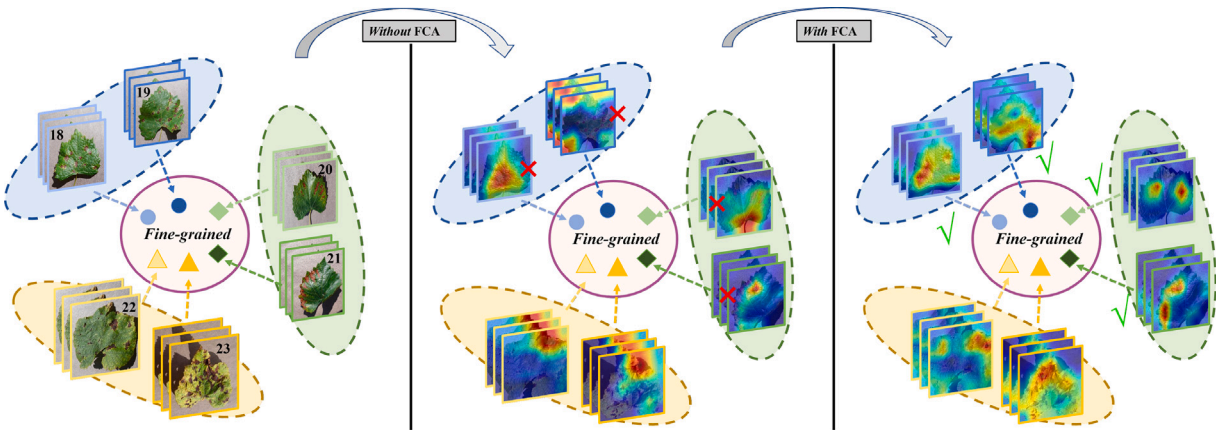
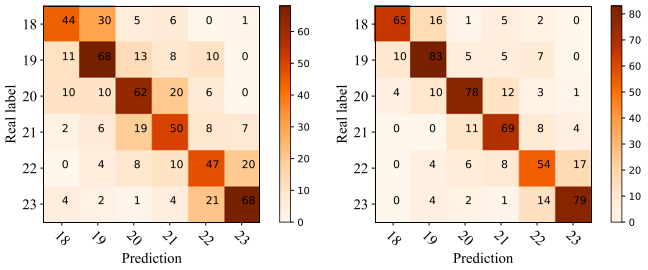


Fig. 12. Visualization of attention heatmaps on FPV. The figure displays six fine-grained disease samples with their attention heatmaps (second column without FCA, third column with FCA). Different shapes represent different diseases (e.g., ○ = *Grape-BlackRotFungus*, △ = *Grape-LeafBlightFungus*). Colors indicate the severity of the diseases (e.g., lighter colors = *general*, darker colors = *severious*). ✗ indicates misclassification, and ✔ indicates correct classification.

Table 5
Comparison of the state-of-the-art few-shot classification algorithms on the FPV dataset. The best results are highlighted in **bold** (mean ± S.D.%). Numbers are in percentage (%).

Methods	FPV	
	5way1shot	5way5shot
Optimization-based		
MAML [67] (2017)	69.96 ± 0.46	82.84 ± 0.40
E^3BM [76] (2020)	78.02 ± 0.42	88.34 ± 0.40
Metric-based		
Relation N [15] (2018)	74.00 ± 0.42	85.86 ± 0.36
Baseline [79] (2019)	71.96 ± 0.37	87.25 ± 0.22
Baseline++ [79] (2019)	76.11 ± 0.40	88.73 ± 0.31
Meta-Baseline [73] (2020)	78.25 ± 0.41	88.76 ± 0.26
Neg-Margin [16] (2020)	78.06 ± 0.46	88.48 ± 0.44
FEAT [80] (2020)	76.25 ± 0.41	88.02 ± 0.24
BML [81] (2021)	77.21 ± 0.63	89.33 ± 0.29
DeepEMD [82] (2020)	76.69 ± 0.47	87.92 ± 0.34
DeepBDC [9] (2022)	79.00 ± 0.52	89.45 ± 0.26
SetFeat12 [84] (2022)	79.21 ± 0.50	89.07 ± 0.28
DeepEMD v2 [46] (2022)	78.86 ± 0.46	89.20 ± 0.28
STANet [10] (2023)	79.14 ± 0.43	89.13 ± 0.30
MB+QCDM [12] (2024)	80.06 ± 0.51	90.06 ± 0.17
Ours	**81.86 ± 0.32**	**91.18 ± 0.26**



(a) Confusion matrix with baseline (b) Confusion matrix with our model

Fig. 13. The confusion matrix for the baseline and ours. The numbers 18, 19, 20, 21, 22, and 23 on the axes represent six fine-grained disease categories of grapes. Each column represents the predicted results, and each row corresponds to the real labels. The six fine-grained categories shown align with the sample categories in Fig. 12.

are presented, with the top showing the visualized visual prompts and the bottom showing the transformed input images. All prompts are generated via sparse updates using SA-MAML.

Cross Alignment Visualization and Case Study. To demonstrate FCA's significance, we compare the attention heatmaps before and after applying CA to samples from different datasets. As shown in Fig. 9, we randomly sample images from *mini*-ImageNet and CUB and visualize their attention heatmaps. The third column shows the self-attention heatmap before adding CA, and the fourth column shows the attention heatmap after cross-alignment. For FPV, we select samples from six fine-grained grape disease categories and visualize their attention heatmaps. The first column in Fig. 12 shows the grape disease images, which are traditionally classified into three categories. However, in this study, they are divided into six classes. The second and third columns visualize the attention heatmaps before and after adding FCA, respectively. Observing Figs. 9 and 12, it is evident that FCA can capture precise discriminative features through cross-image alignment on different granularity datasets. In Fig. 12, class pairs 18–19 and 20–21 are typically considered the most challenging pairs due to their very subtle differences, making it difficult even for experienced agricultural experts to distinguish between them. However, our model provides correct results.

Confusion Matrix and t-SNE Visualization. Fig. 13 shows the confusion matrix comparing our method with another baseline. The numbers on the axes represent six different fine-grained disease categories. In our confusion matrix, the classification results for all difficult-to-classify fine-grained categories have been greatly improved. In addition, we visualize the t-SNE results of these features. Specifically, we resample the data from FPV into three different granularity levels: coarse, intermediate (coarse $< x <$ fine), and fine. And then t-SNE is performed on the high-dimensional representations of samples at different granularity levels. For clarity, we present 2D t-SNE results in Fig. 14. It is visually evident that our model has more easily classifiable feature representations at all granularity levels.

Findings of domain-specific few-shot classification. Compared with existing metric-based and optimization-based methods, FMVP achieves significant performance improvement across different settings. This highlights the proposed approach's advantages:

- **(1)** SMVP enables the visual prompt to learn few-shot-adapted initialization, thus greatly enhancing the model's generalization ability in general-grained few-shot scenarios.
- **(2)** FCA explores intra- and inter-image feature relationships, extracting fine-grained cross-image discriminative features. In fine-grained few-shot tasks where inter-class distributions are similar, this model proves to be highly valuable.
- **(3)** Our model excels in challenging domains with limited data availability, such as fine-grained agricultural disease datasets, showing significant performance gains.

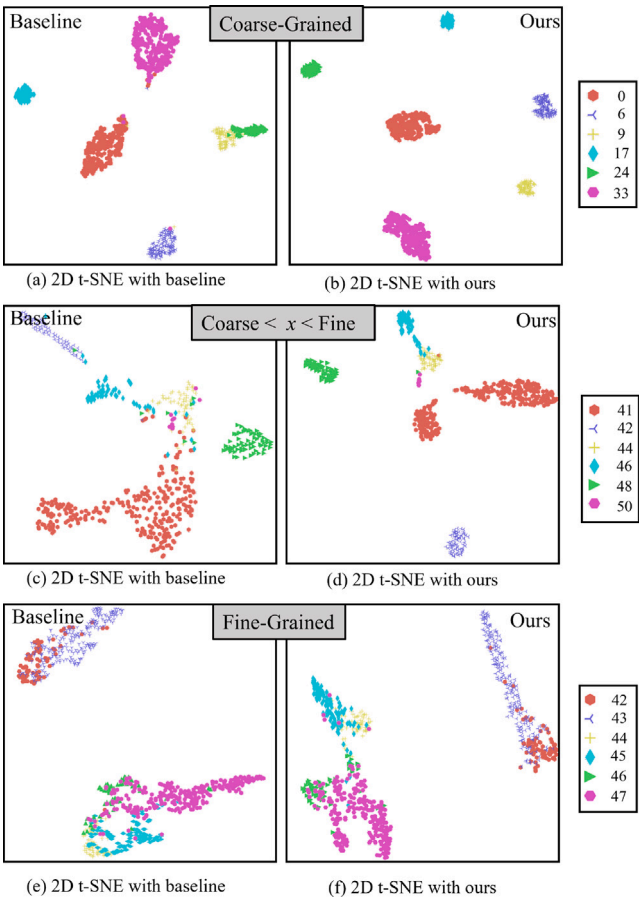


Fig. 14. 2D t-SNE visualization for different granularity levels. Categories range from coarse (species-level) to intermediate (coarse < x < fine, disease-level) and fine (severity-level). Numbers represent class IDs in FPV.

5. Conclusion

For fine-grained few-shot recognition, this paper proposes a novel Fine-grained Meta Visual Prompt tuning (FMVP) framework that combines popular prompt tuning and meta-learning methods. The Sparsity-Aware Meta Visual Prompt tuning (SMVP) sub-module utilizes a unique visual prompt construction method, a novel SA-MAML gradient update paradigm, and a set of carefully designed meta-training tasks to learn a few-shot-adapted prompt initialization, effectively improving generalizability in few-shot scenarios. Additionally, unlike existing self-attention models, we propose a Fine-grained Cross-Alignment (FCA) module to explore discriminative features by extracting fine-grained object-level cross-image semantic relationships. Extensive experiments on classic few-shot datasets and the fine-grained plant disease dataset demonstrate the effectiveness and superiority of our method. Without a doubt, the proposed method is a valuable addition to fine-grained few-shot classification and intelligent agricultural applications.

CRediT authorship contribution statement

Minghui Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hongxun Yao:** Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hongxun Yao reports financial support was provided by National Major Science and Technology Projects of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Science and Technology Major Project, China (2021ZD0110901).

Data availability

No data was used for the research described in the article.

References

[1] P. Ganesan, S.K. Jagatheesaperumal, M.M. Hassan, F. Pupo, G. Fortino, Few-shot image classification using graph neural network with fine-grained feature descriptors, Neurocomputing 610 (2024) 128448.
[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[3] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, Maxvit: Multi-axis vision transformer, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV, Springer, 2022, pp. 459–479.
[4] J. Kang, W. Jia, X. He, Toward extracting and exploiting generalizable knowledge of deep 2D transformations in computer vision, Neurocomputing 562 (2023) 126882.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[7] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[8] S. Wu, H. Luo, X. Lin, TNPNet: An approach to few-shot open-set recognition via contextual transductive learning, Neurocomputing (2024) 129276.
[9] J. Xie, F. Long, J. Lv, Q. Wang, P. Li, Joint distribution matters: Deep Brownian distance covariance for few-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7972–7981.
[10] J. Lai, S. Yang, W. Wu, T. Wu, G. Jiang, X. Wang, J. Liu, B.-B. Gao, W. Zhang, Y. Xie, et al., SpatialFormer: Semantic and target aware attentions for few-shot learning, 2023, arXiv preprint arXiv:2303.09281.
[11] B. Shi, W. Li, J. Huo, P. Zhu, L. Wang, Y. Gao, Global-and local-aware feature augmentation with semantic orthogonality for few-shot image classification, Pattern Recognit. 142 (2023) 109702.
[12] W. Wu, Y. Shao, C. Gao, J.-H. Xue, N. Sang, Query-centric distance modulator for few-shot classification, Pattern Recognit. (2024) 110380.
[13] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. 30 (2017).
[14] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Adv. Neural Inf. Process. Syst. 29 (2016).
[15] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[16] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, H. Hu, Negative margin matters: Understanding margin in few-shot classification, in: European Conference on Computer Vision, Springer, 2020, pp. 438–455.
[17] F. Zhou, L. Zhang, W. Wei, Meta-generating deep attentive metric for few-shot classification, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 6863–6873.
[18] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[19] J. Li, X. He, L. Wei, L. Qian, L. Zhu, L. Xie, Y. Zhuang, Q. Tian, S. Tang, Fine-grained semantically aligned vision-language pre-training, Adv. Neural Inf. Process. Syst. 35 (2022) 7290–7303.

[20] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual prompt tuning, in: European Conference on Computer Vision, Springer, 2022, pp. 709–727.
[21] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, 2021, arXiv preprint arXiv:2104.08691.
[22] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16816–16825.
[23] B. Zhu, Y. Niu, Y. Han, Y. Wu, H. Zhang, Prompt-aligned gradient for prompt tuning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15659–15669.
[24] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, 2021, arXiv preprint arXiv:2101.00190.
[25] A. Li, L. Zhuang, S. Fan, S. Wang, Learning common and specific visual prompts for domain generalization, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 4260–4275.
[26] H. Bahng, A. Jahanian, S. Sankaranarayanan, P. Isola, Visual prompting: Modifying pixel space to adapt pre-trained models, 2022, p. 7, arXiv preprint arXiv:2203.17274 2 (3).
[27] H. Huang, J. Zhang, J. Zhang, J. Xu, Q. Wu, Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification, IEEE Trans. Multimed. 23 (2020) 1666–1680.
[28] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, J. Luo, Revisiting local descriptor based image-to-class measure for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7260–7268.
[29] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, J. Luo, Distribution consistency based covariance metric networks for few-shot learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8642–8649, 01.
[30] L. Tang, D. Wertheimer, B. Hariharan, Revisiting pose-normalization for fine-grained few-shot recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14352–14361.
[31] Y. Zhu, C. Liu, S. Jiang, et al., Multi-attention meta learning for few-shot fine-grained image recognition, in: IJCAI, 2020, pp. 1090–1096.
[32] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, J.-H. Xue, Locally-enriched cross-reconstruction for few-shot fine-grained image classification, IEEE Trans. Circuits Syst. Video Technol. (2023).
[33] R. Ji, J. Li, L. Zhang, J. Liu, Y. Wu, Dual transformer with multi-grained assembly for fine-grained visual classification, IEEE Trans. Circuits Syst. Video Technol. (2023).
[34] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[35] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, D. Tao, Collect and select: Semantic alignment metric learning for few-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8460–8469.
[36] D. Wertheimer, B. Hariharan, Few-shot learning with localization in realistic settings, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6558–6567.
[37] S. Yan, S. Zhang, X. He, et al., A dual attention network with semantic embedding for few-shot learning, in: AAAI, vol. 33, 2019, pp. 9079–9086.
[38] A. Haridasan, J. Thomas, E.D. Raj, Deep learning system for paddy plant disease detection and classification, Environ. Monit. Assess. 195 (1) (2023) 120.
[39] M.G. Selvaraj, A. Vergara, H. Ruiz, N. Safari, S. Elayabalan, W. Ocimati, G. Blomme, AI-powered banana diseases and pest detection, Plant Methods 15 (1) (2019) 1–11.
[40] M. Brahimi, K. Boukhalfa, A. Moussaoui, Deep learning for tomato diseases: classification and symptoms visualization, Appl. Artif. Intell. 31 (4) (2017) 299–315.
[41] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[43] A. Chakraborty, D. Kumer, K. Deeba, Plant leaf disease recognition using fastai image classification, in: 2021 5th International Conference on Computing Methodologies and Communication, ICCMC, IEEE, 2021, pp. 1624–1630.
[44] M.N. Abdal, K. Islam, M.H.K. Oshie, M.A. Haque, A CNN based model for plant disease classification using transfer learning, in: 2023 26th International Conference on Computer and Information Technology, ICCIT, IEEE, 2023, pp. 1–6.
[45] S. Jagadeesan, E. Deepakraj, V. Ramalingam, I. Venkatachalam, M. Vivekanandan, R. Manjula, An efficient detection and classification of plant diseases using deep learning approach, in: 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques, EASCT, IEEE, 2023, pp. 1–6.
[46] C. Zhang, Y. Cai, G. Lin, C. Shen, Deepemd: Differentiable earth mover's distance for few-shot learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (5) (2022) 5632–5648.
[47] C. Liu, Y. Fu, C. Xu, S. Yang, J. Li, C. Wang, L. Zhang, Learning a few-shot embedding model with contrastive learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 8635–8643, 10.
[48] H. Li, L. Li, Y. Huang, N. Li, Y. Zhang, An adaptive plug-and-play network for few-shot learning, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[49] M. Hu, H. Chang, Z. Guo, B. Ma, S. Shan, X. Chen, Understanding few-shot learning: Measuring task relatedness and adaptation difficulty via attributes, Adv. Neural Inf. Process. Syst. 36 (2024).
[50] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.
[51] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[52] H. Bahng, A. Jahanian, S. Sankaranarayanan, P. Isola, Exploring visual prompts for adapting large-scale models, 2022, arXiv preprint arXiv:2203.17274.
[53] Z. Zheng, X. Yue, K. Wang, Y. You, Prompt vision transformer for domain generalization, 2022, arXiv preprint arXiv:2208.08914.
[54] Y. Gao, X. Shi, Y. Zhu, H. Wang, Z. Tang, X. Zhou, M. Li, D.N. Metaxas, Visual prompt tuning for test-time domain adaptation, 2022, arXiv preprint arXiv:2210.04831.
[55] K. Sohn, H. Chang, J. Lezama, L. Polania, H. Zhang, Y. Hao, I. Essa, L. Jiang, Visual prompt tuning for generative transfer learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19840–19851.
[56] J. Wu, X. Li, C. Wei, H. Wang, A. Yuille, Y. Zhou, C. Xie, Unleashing the power of visual prompting at the pixel level, 2022, arXiv preprint arXiv:2212.10556.
[57] A. Chen, Y. Yao, P.-Y. Chen, Y. Zhang, S. Liu, Understanding and improving visual prompting: A label-mapping perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19133–19143.
[58] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4438–4446.
[59] X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, Q. Li, Few-shot learning for domain-specific fine-grained image classification, IEEE Trans. Ind. Electron. 68 (4) (2020) 3588–3598.
[60] X.-S. Wei, J.-H. Luo, J. Wu, Z.-H. Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, IEEE Trans. Image Process. 26 (6) (2017) 2868–2881.
[61] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, Transfg: A transformer architecture for fine-grained recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 852–860, 1.
[62] Z.-C. Zhang, Z.-D. Chen, Y. Wang, X. Luo, X.-S. Xu, ViT-FOD: A vision transformer based fine-grained object discriminator, 2022, arXiv preprint arXiv:2203.12816.
[63] X.-S. Wei, P. Wang, L. Liu, C. Shen, J. Wu, Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples, IEEE Trans. Image Process. 28 (12) (2019) 6116–6125.
[64] Y. Lifchitz, Y. Avrithis, S. Picard, Few-shot few-shot learning and the role of spatial attention, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 2693–2700.
[65] H. Wang, S. Jie, Z. Deng, Focus your attention when few-shot classification, Adv. Neural Inf. Process. Syst. 36 (2024).
[66] L. Zhu, Y. Yang, Compound memory networks for few-shot video classification, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 751–766.
[67] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.
[68] J. Chen, W. Yuan, S. Chen, Z. Hu, P. Li, Evo-MAML: Meta-learning with evolving gradient, Electronics 12 (18) (2023) 3865.
[69] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[70] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[71] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200–2011 dataset, California Institute of Technology, 2011.
[72] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, 2016.
[73] Y. Chen, X. Wang, Z. Liu, H. Xu, T. Darrell, A new meta-baseline for few-shot learning, 2020.
[74] Y. Du, J. Shen, X. Zhen, C.G. Snoek, EMO: Episodic memory optimization for few-shot meta-learning, 2023, arXiv preprint arXiv:2306.05189.

[75] S. Lee, W. Moon, H.S. Seong, J.-P. Heo, Task-oriented channel attention for fine-grained few-shot classification, 2023, arXiv preprint arXiv:2308.00093.
[76] Y. Liu, B. Schiele, Q. Sun, An ensemble of epoch-wise empirical bayes for few-shot learning, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, Springer, 2020, pp. 404–421.
[77] S.-J. Park, S. Han, J.-W. Baek, I. Kim, J. Song, H.B. Lee, J.-J. Han, S.J. Hwang, Meta variance transfer: Learning to augment from the others, in: International Conference on Machine Learning, PMLR, 2020, pp. 7510–7520.
[78] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, L. Sigal, Multi-level semantic feature augmentation for one-shot learning, IEEE Trans. Image Process. 28 (9) (2019) 4594–4605.
[79] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C.F. Wang, J.-B. Huang, A closer look at few-shot classification, 2019, arXiv preprint arXiv:1904.04232.
[80] H.-J. Ye, H. Hu, D.-C. Zhan, F. Sha, Few-shot learning via embedding adaptation with set-to-set functions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8808–8817.
[81] Z. Zhou, X. Qiu, J. Xie, J. Wu, C. Zhang, Binocular mutual learning for improving few-shot classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8402–8411.
[82] C. Zhang, Y. Cai, G. Lin, C. Shen, Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12203–12213.
[83] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, X. He, Learning to affiliate: Mutual centralized learning for few-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14411–14420.
[84] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, C. Gagné, Matching feature sets for few-shot image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9014–9024.
[85] S. Qiu, W. Yang, M. Yang, Hybrid feature collaborative reconstruction network for few-shot fine-grained image classification, 2024, arXiv preprint arXiv:2407.02123.

Minghui Li received the B.S. degree from Civil Aviation University of China in 2018 and the M.S. degree from Sun Yat-sen University in 2021. He is currently pursuing the Ph.D. degree with the Faculty of Computing, Harbin Institute of Technology, Harbin, China. His research interests include computer vision and deep learning, especially focusing on few-shot learning, transfer learning, and domain adaptation.

Hongxun Yao received the B.S. and M.S. degrees in computer science from Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and 1990, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology, in 2003. Currently, she is a professor at the Faculty of Harbin Institute of Technology, a recipient of the Ministry of Education Fund for"the New Century Excellent Talent" in China in 2005, and won the honor title of"enjoy special government allowances expert" in Heilongjiang Province, China. Prof. Yao has mainly researched computer vision intelligence, multimedia data analysis and understanding, affective computing, etc. She has been the executive director of China Society of Image and Graphics, and the director of the Affective Computing and Understanding Special Committee in CSIG. She has published more than 200 papers and achieved one Best Paper Award from ICIMCS 2016. She also has won 1 First Prize and 2 Second Prizes of the Provincial Natural Science and Technology Award. She has published 6 books and owns 16 national invention patents.