Focus nuance and towards diversity : Exploring domain-specific fine-grained few-shot recognition

Minghui Li[1], Hongxun Yao[1*†] and Yong Wang[2†]

[1*]Visual Intelligence Laboratory, Harbin Institute of Technology, Harbin, 150001, China.
[2]MICRO Lab, Sun Yat-sen University, Guangzhou, 510275, China.

*Corresponding author(s). E-mail(s): h.yao@hit.edu.cn;
Contributing authors: 21B903087@stu.hit.edu.cn; wangyong5@mail.sysu.edu.cn;
†These authors contributed equally to this work.

Abstract

In real-world industrial applications, learning to recognize novel visual categories from a few number of samples is challenging and promising. Although some efforts have been made in the academic field for few-shot classification studies, there is still a lack of high-precision fine-grained few-shot classification models in some specific fields, especially in the fine-grained agricultural field. As far as we know, this study is the first work on meta-learning few-shot classification for fine-grained plant disease classification (***specific to disease severity***). We propose a multi-perspective hybrid attention meta-learning model based on a Batch Nuclear-norm constraint. The model explores discriminative features by focusing on key regions, and the hybrid attention module is divided into two sub-modules, soft attention model and patch-hard attention model. The discriminability and diversity constraint module (DDCM) is introduced in the loss function to constrain the Batch Nuclear-norm of the classification matrix, which improves the discriminative properties of the classification model and increases its diversity at the same time. In this paper, a large number of experiments have been carried out on multiple datasets. The experimental results demonstrate that our work has better performance than state-of-the-art (SOTA) models. It can be said that our work is a valuable supplement to the domain-specific industrial application models.

Keywords: fine-grained classification; few-shot learning; visual attention; Batch Nuclear-norm Maximization

1 Introduction

In recent years, we have witnessed significant progress in computer vision thanks to advanced computing devices and large-scale datasets [1, 2]. Deep convolutional neural networks (ConvNets) are able to successfully learn robust feature representations and achieve excellent performance in recognition tasks, which benefits from large-scale labelled training data such as Imagenet [3]. However, such supervised learning methods require a large number of labelled samples and sufficient iterations to train deep learning models and optimize them. As the number of training samples decreases, the performance of deep learning models degrades significantly. In particular, with only one or a few labelled samples, it is difficult to apply the model generalization to a novel category. However, when faced with poorly understood areas, the human brain can draw connections between new concepts and old knowledge to identify novel objects [4]. As a result, the research topic of "recognizing novel visual categories after seeing few labeled samples" has came into being, which is an important attempt to move A.I. towards true "intelligence". The research on this topic is often referred to as "*few-shot learning*".

In practice, it is very difficult to obtain scarce sample datasets in domain-specific scenarios. As an important part of the national strategy, "Smart Agriculture" involves the study of "plant state recognition", which is a typical few-shot learning problem as shown in Figure 1. The rapid detection of plant diseases has serious implications for the safety and sustainability of food production, and is an important indicator for agro-ecosystems. However, due to constraints such as low incidence, high collection costs and difficulty in annotation, very little labeled data on plant status are available. And in particular, accurate annotation of complete plant disease datasets requires a lot of time and expert manpower. It means that collecting large-scale data is both difficult and expensive. As a result, fine-grained few-shot model, which is essential for solving domain-specific problems, has emerged as one of the key topics in computer vision research.

Currently, there are three main approaches in scholarly research to solve few-shot classification problems: model-based learning, optimization-based learning and metric-based learning. These

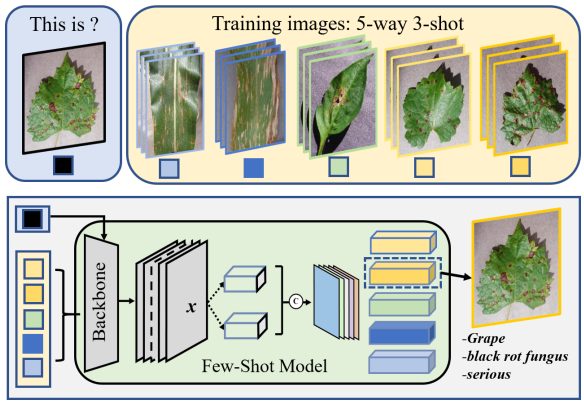


Fig. 1 Brief illustration of fine-grained few-shot classification. Examples of 5-way 3-shot 1-query tasks on fine-grained plant diseases. The current plant disease classification is only accurate to the disease category, and our model can be refined to the disease severity, such as the classification result "*Grape-balck rot fungus-serious*" in the figure.

learning methods differ from previous deep learning models in terms of classification and recognition [3, 5, 6]. Among them, the metric-based learning methods have achieved more advanced research results in recent years. For example, Koch et al. [7] proposed a Siamese network with shared weights for a two-channel convolutional neural network. Different images enter the dual channel separately, and the prediction is obtained by calculating the distance of the output feature vector. The prototype network proposed by Snell et al. [8] maps features to a prototype space. In this space, the classification task can be accomplished by calculating the distance to the prototype representation of each class. The RelationNet proposed by Sung et al. [9] integrates the feature vectors of the support set and the query set, and the relation module compares the two vectors to determine whether they are matching categories. Bin Liu et al. [10] introduces a negative margin loss to the few-shot learning based on metric learning. The experimental results show that the negative margin loss significantly outperforms the softmax loss. Jiangtao X et al. [11] first introduces Brownian Distance Covariance, a probabilistic and statistical-based similarity measure, into deep learning to obtain more accurate similarity by measuring the joint distribution between sample pairs.

Inspired by the above research, we believe that it is important to investigate the detailed relationship between the features of the template image

and the query image for fine-grained few-shot image recognition. Starting with the feature representation, it is essential to investigate a feature extractor that is well trained to accurately represent the category features. Most existing few-shot methods [7–11] focus on finding a suitable distance metric or a meta-learning optimization strategy to compare query images with labelled images. But they barely notice that the key features of fine-grained categories are highly localized and need to be extracted with a relatively local receptive field. For small sample task, focusing on local features can filter out information about distracting objects as well as the background. Limiting the focus of model to where it is more needed is an important means of preventing overfitting. Based on the above observations, we propose a multi-perspective hybrid attention model and a patch-hard attention channel. In this way, we can obtain more discriminatory information by getting focal regions from both soft and hard attention channels. Second, existing methods usually constrain only the discriminability of the classification model in the loss function, while the diversity constraint for rare categories is often ignored. We introduce an additional discriminability and diversity constraint module (DDCM) in the loss function to classify fine-grained hard-to-classify samples by constraining the Batch Nuclear-norm, which improves both the discriminability and diversity of the model by simultaneously constraining the matrix F-norm and the matrix rank. Finally, for domain-specific fine-grained few-shot classification problems, we further perform practical applications in real-world tasks for FinePlantVillage (FPV), a fine-grained dataset, to solve some novel or rare plant disease image classification problems.

To summarize, the main contributions of this paper can be summarized as follows:

- We propose a multi-perspective hybrid attention model (MPHAM) focusing on fine-grained recognition, with the model divided into instance-attention and region-attention subnets internally (soft attention channels and patch-hard attention channels externally). The two sub-network focus on the global information of the image from different perspectives, feeding back a weighted attention feature map. The focal area localization mechanism discovers more local, fine-grained feature information among objects.
- A discriminability and diversity constraint module (DDCM) is introduced to improve the classification diversity of hard-to-classify samples near the decision boundary with high data density. The model is optimized in terms of the discriminability and diversity of the classification matrix by constraining the Batch Nuclear-norm.
- We conduct extensive experiments on the classical few-shot dataset *mini*-Imagenet and *CUB* to prove the validity of our model. Further experiments were conducted on a domain-specific fine-grained few-shot dataset FPV for a real plant disease classification problem. All experiments demonstrate the better performance of the model compared to current SOTA methods.

The rest of the paper is organized as follows. Section 2 summarizes related work on few-shot learning, fine-grained image recognition, and plant disease classification. Section 3 describes the proposed MPHAM and DDCM methods in detail, and then, Section 4 presents the experimental setup, ablation experiments and analysis of experimental results on different datasets. Finally, Section 5 concludes the paper and provides an outlook for future research.

2 Related Work

This section briefly introduces related research areas to define and describe our own proposed methods. First, it is introduced that the current research status of plant disease classification algorithms, and then, two closely related areas: few-shot learning and fine-grained recognition are introduced. Finally, we review related studies and present our proposed solution.

2.1 Plant Disease Classification

Plant Disease Classification. Deep learning has made significant achievements for a wide range of intelligent perception tasks such as classification, detection and segmentation, which has led to its widespread use in agricultural scenarios. Deep learning-based recognition algorithm significantly improves the accuracy of plant pest and disease

detection. In 2018, Ferentinos et al. [12] open-sourced a plant disease dataset of 87848 images for 58 diseases of 25 plant species and achieved 99% classification accuracy using convolutional neural networks (CNN) structures. In 2019, Selvaraj et al. [13] achieved 94.1% accuracy for pest and disease detection of banana using Faster RCNN model. In 2021, Aboneh T et al. [14] developed a wheat disease classification system based on the VGG19 model that achieved 99.38% accuracy in classifying wheat image data collected from the Bishoftu region.

The deep convolutional neural networks mentioned above usually require thousands of labeled examples of each class to achieve good recognition results. However, in practical applications, it is impractical to collect large amounts of labeled samples. Especially for uncommon diseases, often only a few or dozens of disease images can be collected, which cannot reach the number of samples required for deep network training, resulting in insufficient training of the network and poor generalization effect. Although the number of training samples can be reduced to some extent by migration learning, it cannot solve the problem of easy overfitting of the model due to training with a small number of samples. Humans, by contrast, can quickly identify a new category of objects. Inspired by such strong learning ability, researchers have proposed the concept of few-shot learning, which can have the ability to learn from a small number of data samples like humans do.

2.2 Few-shot learning

Convolutional neural networks typically require thousands of labeled samples per class to achieve the required performance. However, it is impractical to collect large amounts of annotated data, especially for domain-specific industrial applications that require specialized knowledge. Some researchers have started to focus on few-shot visual classification studies.

Optimization-based few-shot learning. The parameter optimization-based approach learns a general strategy for fast parameter optimization that finetunes or predicts the learner parameters for each specific few-shot learning task. Santoro et al. [15] trained a cross-task meta-learner that can quickly and accurately update the parameters in the model. Finn et al. [16] trained a model-independent meta-learner and found initial parameters adapted to various tasks with similar distributions, which can be quickly generalized to new tasks with few training samples by setting the initialization parameters obtained from learning. Ravi and Larochelle [17] proposed a meta-learning model based on a long short-term memory network (LSTM) to learn the general initialization and the update rules of classifier parameters. Momin Abbas et al. [18] proposed a sharpness aware Model-Agnostic Meta-Learning (MAML) method using sharpness aware minimization to avoid the loss function from falling into local optimum as much as possible.

Metric-based few-shot learning. The metric learning-based approach is a classification method that measures the similarity between the query image and the support image by learning the embedding function. During the testing period, the nearest neighbor method was used to classify new categories in the embedding space, where samples with the same categories were closer than that with different categories. Snell et al. [8] proposed a prototypical network (Proto-Net) to learn the prototypes of each category and classify them by computing the Euclidean distance between the query image and the prototype in the embedding space. Unlike the Proto-Net which manually selects fixed metrics (e.g., cosine and Euclidean distance), the RelationNet [9] uses a nonlinear comparator for learning and directly compares the metric distance between the query image and the support image in the embedding space. A Meta-Baseline method was proposed by Yinbo Chen et al [19]. The method pretrains the classifier on all base classes and performs meta-learning on a metric-based few-shot classification algorithm.

2.3 Fine-grained image recognition

Fine-grained image classification faces the challenge of small differences among subclasses and large differences among images within classes. It is dedicated to solving the fine-grained analysis problem in image classification. The current mainstream approach is to first locate the most discriminative region in fine-grained images, and then classify it using the local features obtained [20–22]. Multi-attention CNN (MA-CNN) [23] learns fine-grained features better by cross-training two subnetworks. In view of the fact that region detection

and fine-grained feature learning are independent of each other in the existing fine-grained methods, Fu et al. [20] proposed a multiscale recurrent attention convolutional network that recursively learns discriminative region attention and region-based feature representations by mutual reinforcement. To classify fine-grained vision objects in multiple regions, Shen Chen et al. [24] proposed a Context-aware Attentional Pooling (CAP), which can help the model better learn the visual features of each part of the object.

Summarizing the research status in the above fields, we can see that deep Conv-Nets have made remarkable achievements in a wide range of visual tasks [2, 12–14, 20–24]. However, obtaining reliable discriminative representations and good model generalization for fine-grained few-shot image classification is still a rather challenging problem. Different from these methods, this paper incorporates a multi-perspectives hybrid attention model (MPHAM) into the feature extractor based on a few-shot learning approach, so that the feature extractor focuses more on salient regions and avoids losing useful information. Also, a discriminability and diversity constraint module (DDCM) is introduced in the loss function section to better measure the distance between two fine-grained images and to effectively improve the classification diversity.

3 Proposed Method

Problem definition. According to the standard definition of few-shot classification task[8, 9, 25, 26], we divide the plant disease dataset into a base class dataset $\mathcal{D}_{\text{base}} = \{(x_i, y_i)\}_{i=1}^{N}$ and a novel class dataset $\mathcal{D}_{\text{novel}}$, where $\boldsymbol{x}_i$ is an image sample and $\boldsymbol{y}_i$ is the truth label of $\boldsymbol{x}_i$, $\mathcal{D}_{\text{base}} \bigcap \mathcal{D}_{\text{novel}} = \emptyset$, and N denotes the total number of images. To accomplish an C-way K-shot Q-query task, a support set S and a query set Q are partitioned within the dataset:

$$\begin{aligned} \mathcal{S} &= \{(x_i^s, y_i^s)\}_{i=1}^{N_s} \, (N_s = K * C) \\ \mathcal{Q} &= \{(x_j, y_j)\}_{j=1}^{N_q} \end{aligned} \tag{1}$$

where K denotes K images with labels and C denotes the number of novel classes. A few-shot task defined in this way is called a C-way K-shot setting.

Overall framework. Few-shot classification networks usually contain a feature extractor and a predictor. In order to improve the fine-grained classification capability of hard-to-classify samples near the decision boundary with high data density and enhance the feature extraction capability, a novel few-shot classification model is proposed in this paper. Our approach is outlined in Figure 2, which mainly consists of a multi-perspective hybrid attention module (MPHAM) and a predictor for improving discriminability and diversity: the DDCM module. The internal module of MPHAM includes an instance attention module and a region attention module, which focuses on the global information of the image from different perspectives. The external module of MPHAM is divided into soft attention and patch-hard attention channels, which are later docked to the DDCM module for final result prediction and to enhance the discriminability and diversity of classifiers. The whole network is based on convolutional units and measurement function, which can be trained end-to-end, and the details of each component are described below.

3.1 Multi-perspective hybrid attention module

The commonly used feature extraction backbone networks (e.g., ResNet [27], VGG [5], etc.) have achieved excellent results on general classification problems. However, the results of these networks on few-shot classification tasks are not satisfactory. Existing methods are particularly ineffective when faced with fine-grained image recognition problems (e.g., plant disease classification). Therefore, the network with a classical classification backbone as the feature encoder can't meet the requirements of fine-grained image recognition.

In this paper, we integrate the multi-perspective hybrid attention module (MPHAM) into the feature encoder to improve the adaptability of the model and realize the attention fusion of region and channels. Finally, an attention-based feature encoder is constructed. As shown in Figure 2, MPHAM considers the features extracted by the backbone network as prior knowledge of the subsequent attention-based learning network, and then focuses on the global information of the image from different perspectives using instance attention and region attention. The

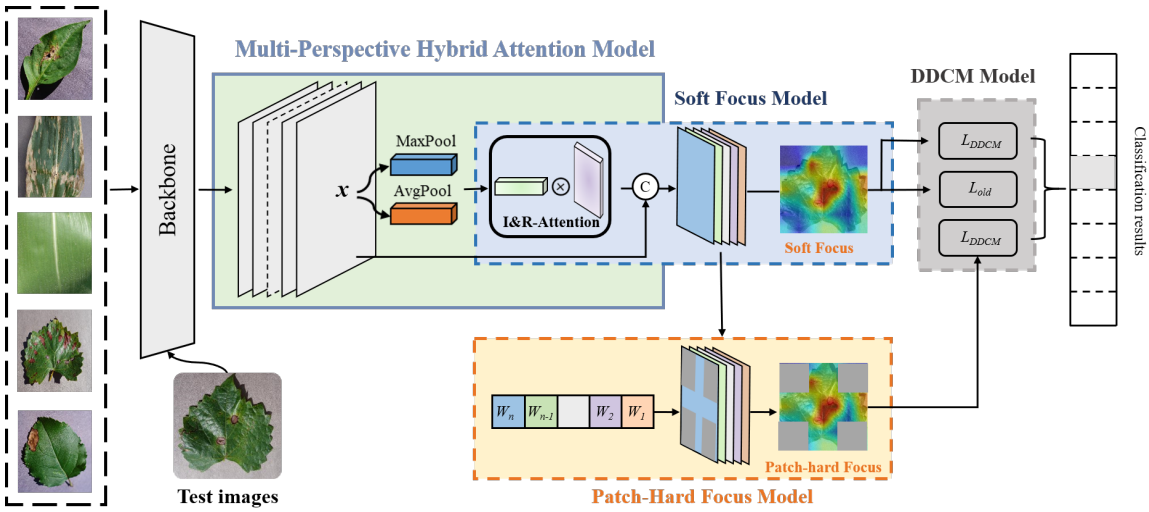


Fig. 2 Illustration of proposed framework on a 5-way 1-shot fine-grained plant disease classification task. We first use the backbone network for feature extraction, and then refine the extracted features by focusing on the discriminative regions with a multi-angle hybrid attention module(MPHAM). In loss function part, we introduce the discriminability and diversity constraint loss (DDCM), which effectively improves the classification accuracy of hard-to-classify samples.

attention module doesn't require separate pre-training for it, simplifying the training process. The detailed structure of MPHAM is shown in Figure 3, which contains two modules: instance attention module and region attention module, and the two attention modules are connected in a tandem manner. It helps the network to extract the dataset feature distribution effectively.

Instance attention module. Instance attention can tell the model which channel is worthy of special attention. We assume that the input feature map of the network is $\boldsymbol{P} \in \mathbb{R}^{C\times H\times W}$. Firstly, global average pooling (GAP) and global maximum pooling (GMP) are executed to aggregate spatial information of a feature map. Then two 1 x 1 x C feature maps are generated: $\boldsymbol{P}^c_{\text{avg}}$ and $\boldsymbol{P}^c_{\text{max}}$. Secondly, these two feature maps are fed into the multi-layer perceptron (MLP) separately, and shares MLP parameters. Then, the two feature maps are added together and the weight coefficients between 0 and 1 are obtained by the sigmoid function σ. Finally, the weight coefficients are then multiplied with the input feature map to get the final weighted feature map $\boldsymbol{\beta}^c$.

$$\begin{aligned}\boldsymbol{\beta}^c &= \sigma(\text{MLP}(\text{Avg Pool}(\boldsymbol{P})) + \text{MLP}(\text{MaxPool}(\boldsymbol{P})))\\ &= \sigma\left(\boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{P}^c_{avg}\right)\right) + \boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{P}^c_{\text{max}}\right)\right)\right)\end{aligned} \tag{2}$$

where σ denotes the sigmoid function. $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ are the parameters to be learned in the MLP network.

For 1-shot task, $\boldsymbol{\beta}^c$ is the final prototype representation. For 5-shot task, the five feature maps are weighted and summed to obtain a prototype representation ξ that summarizes all the information of a certain class.

$$\xi = \sum_{i=1}^{5} \beta_i^c \boldsymbol{X}_i \tag{3}$$

where $\boldsymbol{X} = [x_1, \ldots, x_n]$ is the image feature extracted by feature extractor. $\boldsymbol{X}_i$ represents the feature of the ith image.

Region attention module. We introduce the region attention module to focus on which part of the spatial dimension has more significant features. We aggregate the high-order integration information of a feature map by using the average-pooling and max-pooling operations, and then stitch the two feature maps together in the channel dimension. The processed feature maps then pass through a convolution layer with a convolution kernel of 7 x 7 and reduce to 1 channel, while keeping H and W unchanged. Finally, like the instance attention module, the region weight coefficients are generated by the sigmoid function σ, and then multiplied with the input feature map to obtain the final weighted feature map $\boldsymbol{\beta}^s \in \boldsymbol{R}^{H\times W}$.

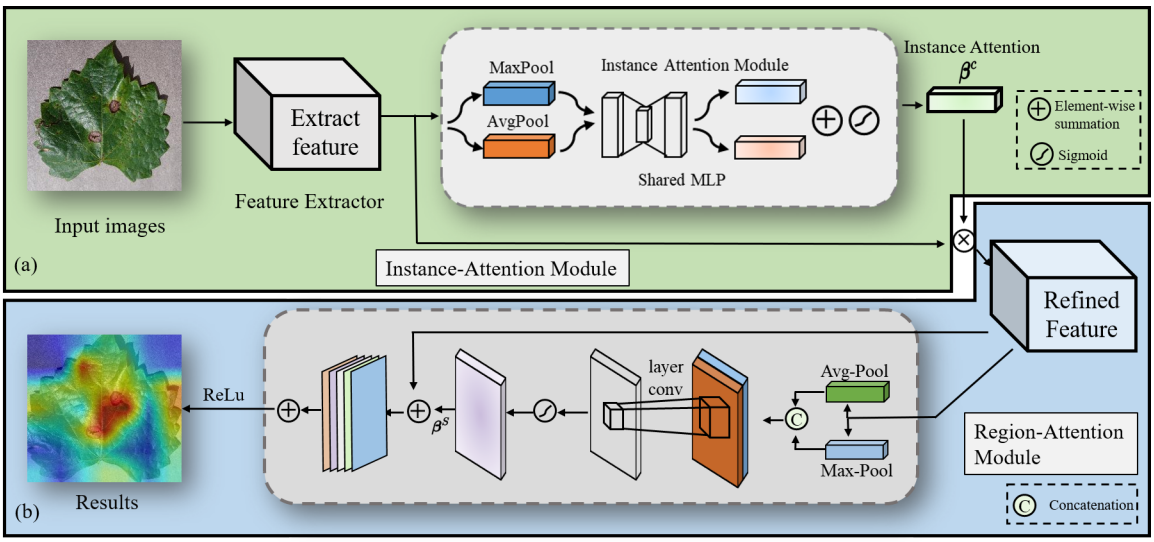


Fig. 3 The proposed multi-perspective attention module (MPHAM). (a) Illustration of the instance attention module (b) Illustration of the region attention module.

$$\begin{aligned}\boldsymbol{\beta}^s &= \sigma\left(([\text{AvgPool}(\boldsymbol{SP});\ \text{MaxPool}(\boldsymbol{SP})]\boldsymbol{W}_2 + b_2)\right) \\ &= \sigma\left(\left(\left[\boldsymbol{SP}^s_{\text{avg}}\ ;\ \boldsymbol{SP}^s_{\text{max}}\right]\boldsymbol{W}_2 + b_2\right)\right)\end{aligned} \tag{4}$$

where SP indicates the input feature map of the spatial attention module, σ denotes the sigmoid function, and $\boldsymbol{W}_2$ and b_2 are the parameters to be learned in the network.

For 1-shot task, $\boldsymbol{\beta}^s \in \boldsymbol{R}^{H\times W}$ is the final prototype representation. For 5-shot task, the five feature maps are weighted and summed to obtain a prototype representation ψ that summarizes all the information of a certain class. $\boldsymbol{X}_i$ represents the feature of the ith image.

$$\psi = \sum_{i=1}^{5} \beta_i^S X_i \tag{5}$$

Patch-hard attention channel. In the soft attention mechanism, all information is re-weighted in an adaptive manner before being aggregated. Compared to the soft attention mechanism, hard attention selects only a portion of the information and processes it further. Hard attention can improve accuracy and learning efficiency by focusing computation on the important parts of an image and ignoring some redundant information. The proposed patch-hard attention in this paper processes only the part of information that is considered to be the most relevant. For the specific implementation, the few patches with the highest average response in the weighted feature map are retained and the lower weighted patches are masked. Then the hard attention weighted feature map is generated.

Specifically, each weighted feature map $\boldsymbol{\beta}^s \in \boldsymbol{R}^{H\times W}$ generated by region attention model is divided into 9 patches evenly. Then the k regions with the highest weight response are retained, and the rest of the patches are masked. Assuming that the weights sum of each patch is $\mathbb{W}_p$, the k patches finally retained can be expressed as $\mathbb{P}_R$:

$$\mathbb{P}_R = \{\text{top}\, k\,(\mathbb{W}_p), k \in (1, ..., 9)\} \tag{6}$$

3.2 Discriminability and diversity constraint module

In migration learning tasks, the target domain often leads to the existence of some indistinguishable similar data near the decision boundary due to the lack of labels. In order to deal with this problem, we analyze the batch classification matrix A and hope to optimize it in terms of discriminability and diversity.

A. Measuring Discriminability with F-norm.

Suppose the prediction matrix of the model for a batch data is $\mathbf{A} \in \mathbb{R}^{B\times C}$, where B and C denote the batch size and the number of categories,

respectively, and satisfies:

$$\sum_{j=1}^{C} \mathrm{A}_{i,j} = 1 \quad \forall\, i \in 1 \ldots B$$
$$\mathrm{A}_{i,j} \geq 0 \quad \forall\, i \in 1 \ldots B, j \in 1 \ldots C \tag{7}$$

For classical classification models with sufficient samples, a good performance classification matrix A can be obtained by training a sufficient number of labeled samples. However, when there are insufficient labels, there are often many ambiguous samples near the decision boundary that are easily misclassified. Existing methods usually optimize the prediction results for unlabeled samples by increasing discriminability [28]. It has been shown in [29] that maximizing the F-norm of A can constrain the model prediction discriminability. The optimization objective can be expressed as:

$$\|\mathrm{A}\|_F = \sqrt{\sum_{i=1}^{B}\sum_{j=1}^{C}|\mathrm{A}_{i,j}|^2} \tag{8}$$

$\|\mathrm{A}\|_F$ is the F-norm of classification matrix A, which has strict opposite monotonicity with the entropy $H(\mathrm{A})$. The classification discriminability can be improved by maximizing $\|\mathrm{A}\|_F$ in the loss function.

B. Measuring Diversity with Matrix Rank.

It is normal for some categories to have a majority of samples in a randomly selected batch of samples, while other categories contain fewer or even no samples. In this case, the model trained using entropy minimization or F-norm maximization tends to classify samples near the decision boundary as the majority class. The continuous convergence to most categories reduces the diversity of predictions, which is detrimental to the overall prediction accuracy. To improve classification accuracy, unlike other methods [30–33], we aim to maintain predictive diversity by analyzing the batch output matrix A.

Since the data of each batch is randomly sampled, the expectation of category in each batch should be stable. Noting this property, we can constrain the diversity by constraining rank(A) (the rank of the classification matrix A) to be maximal. It prevents the predictions of the model from collapsing to the majority category. But the rank of the matrix is difficult to optimize directly, which is an NP-Hard problem.

C. Batch Nuclear-norm Maximization.

In order to be able to constrain both discriminability and diversity, we find that there is a relationship between the nuclear-norm $\|\mathrm{A}\|_*$ and the rank of matrix $\|\mathrm{A}\|_F$. In [34–36] , this relationship can be expressed as:

$$\frac{1}{\sqrt{D}}\|\ \mathrm{A}\|_* \leq \|\mathrm{A}\|_F \leq \|\mathrm{A}\|_* \leq \sqrt{D} \cdot \|\mathrm{A}\|_F \tag{9}$$

where $D = min(B, C)$. This shows that $\|\mathrm{A}\|_*$ and $\|\mathrm{A}\|_F$ can be bounded to each other. Then it is natural that

$$\|\mathrm{A}\|_* \leq \sqrt{D} \cdot \|\mathrm{A}\|_F \leq \sqrt{D \cdot B} \tag{10}$$

Therefore, maximizing $\|\mathrm{A}\|_*$ can force $\|\mathrm{A}\|_F$ to maximize, which in turn makes the model predictions both discriminative and diverse.

Figure 4 shows the classification comparison between Entropy Minimization (EntMin) and DDCM when processing the same classification matrix A. Compared with the previous EntMin, this classification matrix is constrained by DDCM with batch nuclear-norm maximization to obtain more accurate classification results, which improves the classification diversity.

3.3 Loss Function Fusion

The proposed framework shows that the attention channel is divided into a patch-hard attention channel and a soft attention channel. The patch-hard attention channel aims to improve discriminability by concentrating on a few representative regions, which are later constrained by DDCM. The soft attention channel covers the full feature map and uses both DDCM loss and cross-entropy loss as the final optimization loss function for this channel.

In the task, we are given labeled domain $\mathcal{D}_L$ and unlabeled domain $\mathcal{D}_U$. Classification results are obtained by the deep network $G(x_i)$. For a randomly sampled batch B_L example $\{X^L, Y^L\}$ on a labeled data set, where X^L denotes the sample with label, Y^L denotes the label corresponding to the sample. The classification loss on $\mathcal{D}_L$ can be calculated as:

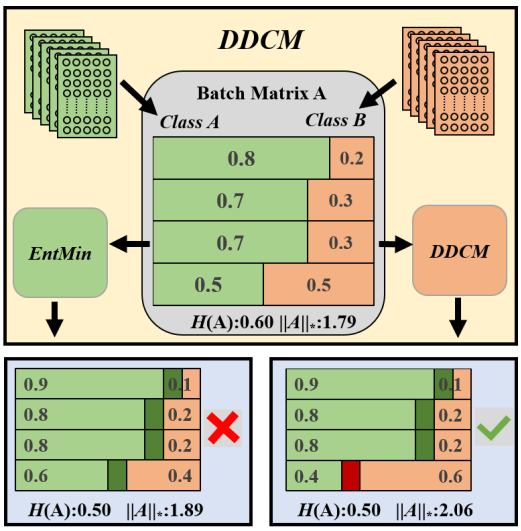


Fig. 4 Comparison of classification effects of DDCM and Entropy Minimization (EntMin). In the figure, the two categories of the examples are *Class 01* and *Class 02*. Batch Matrix A is the original classification matrix, and two new classification matrices are generated by DDCM and EntMin. The dark green (red) represents the increase of green (red) variable. $H(\mathrm{A})$ represents the entropy value and $\|\mathrm{A}\|_*$ represents the value of nuclear-norm.

$$\mathcal{L}_{cls} = \frac{1}{B_L} \left\| Y^L \log \left(G \left(X^L \right) \right) \right\|_1 \tag{11}$$

On the unlabeled domain, the DDCM loss is applied to the classification matrix. For a randomly sampled batch B_U example $\left\{X^U, Y^U\right\}$, the classification matrix on $\mathcal{D}_U$ can be denoted as $G\left(X^U\right)$. The loss function of DDCM can be expressed as:

$$\mathcal{L}_{ddcm} = -\frac{1}{B_U} \left\| G \left(X^U \right) \right\|_* \tag{12}$$

To train the network, for the soft attention channel we optimize both the classification loss and the DDCM loss, $\mathcal{L}_{cls}$ and $\mathcal{L}_{ddcm}$ can be optimized simultaneously and combined with the parameter λ as follows

$$\mathcal{L}_{soft} = \frac{1}{B_L} \left\| Y^L \log \left(G \left(X^L \right) \right) \right\|_1 - \frac{\lambda}{B_U} \left\| G \left(X^U \right) \right\|_* \tag{13}$$

For the patch-hard attention channel, we only optimize the DDCM loss to increase its discriminability, whose loss function is expressed as:

$$\mathcal{L}_{patch-hard} = -\frac{1}{B_U} \left\| G \left(X^U \right) \right\|_* \tag{14}$$

The overall loss is expressed as:

$$\mathcal{L}_{all} = \mathcal{L}_{soft} + \mathcal{L}_{patch-hard} \tag{15}$$

4 Experimental result and analysis

4.1 Experiments

This paper focuses on the fine-grained few-shot classification task. In order to reasonably evaluate the effect of the prosed model, several different experiments are designed on few-shot classification datasets such as the *mini*-ImageNet [37] and Caltech-UCSD-Birds [38] datasets. Then it is applied to a modified fine-grained plant disease classification dataset: *FPV*. *FPV* is created based on PlantVillage (PV) and contains 10 species with 27 diseases. Each image is marked with a label accurate to the severity of the disease, which format is "plant-disease-severity", and is subdivided into 61 species. This dataset is a more fine-grained dataset than the original PV dataset, as shown in Figure 5.

In the ablation experiments, this paper verifies the effect of the combination of HPHAM and DDCM, through a series of combinations of experimental modules. First, for the internal submodules of HPHAM: instance-attention and region-attention, ablation experiments are performed to verify the effectiveness of the two modules when connected in series. Second, patch-hard attention channel is treated as an attention module independent of MPHAM to carry out combinatorial experiments, verifying its superiority. Finally, DDCM is also integrated into the experiment as a configuration module, and the simultaneous use of each module achieves SOTA effect.

In the experimental results and analysis section, with the n-way k-shot experimental setup, experiments are conducted under uniform parameter settings to compare the superiority of the proposed method over SOTA methods. In each experimental evaluation, C categories (C-way) are randomly selected in a given test domain, and then N images from each category are selected as the support set (N-shot, a total of C*N images) and M images as the query set (a total of C*M images). 5-way 1-shot and 5-way 5-shot are selected to evaluate the model. The division of base class and

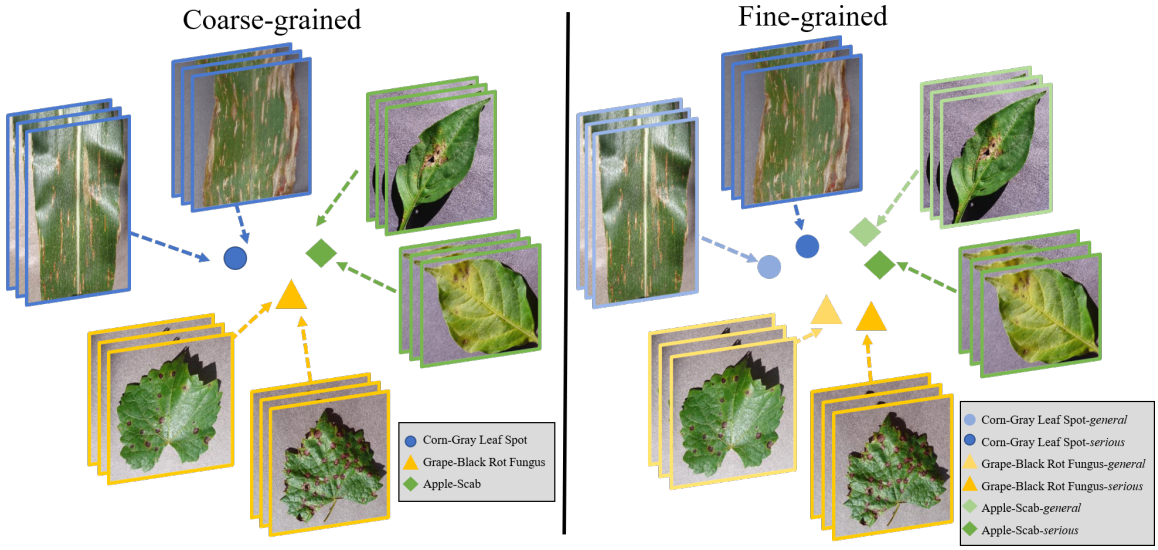


Fig. 5 Example of classification tasks for coarse-grained *PV* dataset and fine-grained *FPV* dataset. (a) The left side of the figure shows the coarse-grained categories in the format "plant-disease", and (b) the right side shows the fine-grained categories in the format "plant-disease-severity". The left and right images are identical, but the categories are completely different at different fine-grained levels.

novel class in *CUB* and *mini*-Imagenet is the same as that in [8, 17, 25]. In *FPV*, 30 classes are selected as base classes, and the rest are classified as novel classes. The model evaluation metrics are aligned with the mainstream metrics [8, 10, 19].

4.2 Implementation Details

ResNet12 [27] is serving as the feature extractor backbone. The original support images are uniformly adjusted to 84×84 and sent to the feature extractor to form robust feature vectors. Momentum and learning rate are set to 0.9 and 0.001 respectively. The model was trained over 5000 iterations. All experiments are conducted by PyTorch on NVIDIA 2080 Ti GPU. To reasonably measure the model performance, we use the top-1 accuracy as an evaluation metric.

4.3 Ablation Studies

During the ablation experiment, ResNet12 is used as the basic backbone of the feature encoder. The modules to be configured include the MPHAM, the patch-hard attention channel and DDCM. It should be reminded that patch-hard attention channel is only a sub-module of MPHAM. But when conducting ablation experiments, they are treated as two "separate" modules.

As shown in Table 1, in the classification task, if the module is added then mark ($\surd$) the appropriate place in the table, otherwise leave it blank. Several marks ($\surd$) means that several modules are used at the same time. Since the patch-hard attention channel is dependent on the hybrid attention module, it cannot be used as a separate module without MPHAM. Therefore, when the modules are combined, the patch-hard attention channel does not appear alone. The baseline values in the table are the results of the Meta-Baseline [19].

The influence of Instance/Region attention module. The hybrid attention module consists of two parts: the instance attention sub-module and the region attention sub-module. Table 1 presents the classification results in the test set. The model using the individual attention module alone performs better than the one without the hybrid attention module. It demonstrates the effectiveness of both attention submodules. The instance attention module can adaptively fuse high-order information to generate better class prototypes, while the region attention module can adaptively select perceptual fields to extract more effective image features. Meanwhile, the combination of the instance attention module and the region attention module (i.e., hybrid attention) utilizes the advantages of both high-order integration and one-order location. It can be concluded that the

Table 1 Few-shot results (5-way 1-shot) of proposed model with different settings of attention mechanism. Best results are displayed in boldface. Numbers are in percentage.

Baseline	Instance-A	Region-A	mini-ImageNet	*CUB*	*FPV*
√			67.04 ± 0.63	77.21 ± 0.63	76.25 ± 0.41
√	√		67.25 ± 0.42	78.31 ± 0.58	78.77 ± 0.35
√		√	67.12 ± 0.28	78.22 ± 0.60	78.51 ± 0.46
√	√	√	**68.25 ± 0.44**	**78.71 ± 0.65**	**79.02 ± 0.35**

Table 2 Ablation study of 5-way 1-shot on few-shot classification. Best results are displayed in boldface. Numbers are in percentage.

Baseline	HAM	Patch-HA	DDCM	*mini*-ImageNet	*CUB*	*FPV*
√				67.04 ± 0.63	77.21 ± 0.63	76.25 ± 0.41
√	√			68.25 ± 0.44	78.71 ± 0.65	79.02 ± 0.35
√			√	67.32 ± 0.34	78.25 ± 0.65	78.77 ± 0.46
√	√	√		68.31 ± 0.31	78.77 ± 0.36	79.21 ± 0.38
√	√		√	68.48 ± 0.26	79.07 ± 0.46	79.65 ± 0.46
√	√	√	√	**68.75 ± 0.30**	**79.48 ± 0.47**	**79.84 ± 0.33**

combination of these two modules achieves the best performance.

The influence of MPHAM. The experimental results based on the multi-perspective hybrid attention module demonstrate that the attention module has significantly improved the classification results on the fine-grained dataset *CUB* and the fine-grained disease classification dataset *FPV*, with the improvement of classification accuracy by about 1.5% and 2.7%, respectively. Multiple experimental results on all three datasets show that the MPHAM benefits both fine-grained classification task and coarse-grained classification task. Figure 8 shows the focusing effect of attention module.

The influence of Patch-hard attention module. The patch-hard attention module can only appear based on the hybrid attention module, so it does not appear alone. Based on Experiments 2, 4, 5 and 6 in Table 2, it is observed that after adding patch-hard attention channel to HAM, the model has better performance on fine-grained classification datasets. Experiments 5 and 6 show that the addition of the patch-hard attention channel improves the model accuracy by 0.27% (*mini*-Imagenet), 0.41% (*CUB*) and 0.19% (*FPV*) compared to the case with only the hybrid attention module.

The influence of DDCM. The discriminability and diversity constraint module (DDCM) appears in the model as a constraint on the original loss function. It can be seen that DDCM already brings some improvement to the model performance when it appears alone in Experiment 3. And when it is combined with the hybrid attention module and patch hard attention channel in Experiment 6, it shows a greater performance improvement and achieves the best accuracy.

In summary, a series of ablation experiments on the proposed method in this paper verified the positive impact of each module on the performance of fine-grained classification.

4.4 Comparison with state-of-the-arts

By ablation experiments and analysis, it has been validated that each component in the model has performance advantages.

Next, to demonstrate the optimization capability of the model in this paper, a series of experiments are performed on the *mini*-ImageNet [37] and *CUB* [38]. 5-way 1-shot and 5-way 5-shot are selected to evaluate the model. Table 3 shows the comparison results of this method with the most advanced method under each dataset.

Table 3 Comparison of the state-of-the-art few-shot classification algorithms on the *mini*-ImageNet and *CUB* dataset. The best results are highlighted in bold.

Methods	miniImageNet		*CUB*	
	5way1shot	5way5shot	5way1shot	5way5shot
Optimization-based				
MAML [16]	57.40 ± 0.47	72.42 ± 0.65	70.44 ± 0.55	85.50 ± 0.33
Metric-based				
Matching N [25]	59.30 ± 0.44	72.63 ± 0.36	78.33 ± 0.45	88.98 ± 0.26
Relation N [9]	54.12 ± 0.46	71.31 ± 0.37	73.22 ± 0.48	86.94 ± 0.28
Prototypical N [8]	56.13 ± 0.45	75.70 ± 0.33	74.35 ± 0.48	88.50 ± 0.25
Baseline [39]	60.00 ± 0.44	80.55 ± 0.31	71.85 ± 0.46	88.09 ± 0.25
Baseline++ [39]	63.25 ± 0.44	81.67 ± 0.30	75.25 ± 0.45	89.85 ± 0.23
Meta-Baseline [19]	64.17 ± 0.45	81.41 ± 0.31	78.16 ± 0.43	90.04 ± 0.23
Neg-Margin [10]	61.70 ± 0.46	78.03 ± 0.33	78.14 ± 0.46	90.00 ± 0.24
FEAT [40]	66.78 ± 0.20	82.05 ± 0.14	77.53 ± 0.83	89.79 ± 0.28
BML [41]	67.04 ± 0.63	83.63 ± 0.29	77.21 ± 0.63	90.45 ± 0.36
DeepEMD [42]	65.91 ± 0.82	82.41 ± 0.56	75.65 ± 0.63	88.69 ± 0.50
DeepBDC [11]	67.83 ± 0.43	84.45 ± 0.29	79.01 ± 0.42	90.42 ± 0.17
Ours	**68.75 ± 0.44**	**85.25 ± 0.32**	**79.48 ± 0.47**	**91.05 ± 0.26**

Table 4 Few-shot results with different settings of backbones. Best results are displayed in boldface. Numbers are in percentage.

Methods	Backbones	mini-ImageNet		*CUB*		*FPV*	
		5way1shot	5way5shot	5way1shot	5way5shot	5way1shot	5way5shot
Baseline [39]	conv4	46.06 ± 0.39	65.83 ± 0.35	47.73 ± 0.41	68.77 ± 0.38	46.24 ± 0.30	66.50 ± 0.28
	resnet12	60.00 ± 0.44	80.55 ± 0.31	71.85 ± 0.46	88.09 ± 0.25	71.96 ± 0.37	87.25 ± 0.22
Baseline++ [39]	conv4	51.16 ± 0.43	67.99 ± 0.36	62.01 ± 0.49	77.72 ± 0.36	56.61 ± 0.44	76.60 ± 0.21
	resnet12	63.25 ± 0.44	81.67 ± 0.30	75.25 ± 0.45	89.85 ± 0.23	76.11 ± 0.40	88.73 ± 0.31
Meta-Baseline [19]	conv4	51.35 ± 0.42	66.99 ± 0.37	58.98 ± 0.47	75.77 ± 0.37	56.25 ± 0.45	75.02 ± 0.34
	resnet12	64.17 ± 0.45	81.41 ± 0.31	78.16 ± 0.43	90.04 ± 0.23	78.25 ± 0.41	90.00 ± 0.26
Ours	conv4	51.25 ± 0.40	68.05 ± 0.33	62.15 ± 0.44	78.12 ± 0.30	57.66 ± 0.31	77.28 ± 0.32
	resnet12	**68.75 ± 0.44**	**85.25 ± 0.32**	**79.48 ± 0.47**	**91.05 ± 0.26**	**79.84 ± 0.33**	**90.25 ± 0.23**

In the study of few-shot learning, the same backbone network is usually used for effect comparison. The two commonly used backbone networks are Conv4 and ResNet12 [8, 16, 25]. It can be seen from Table 4 that the proposed model has a more significant performance improvement on both the shallow backbone network Conv4 and the deeper backbone network ResNet12. In order to obtain higher classification accuracy and fairly compare detection results, all the following experiments are based on ResNet12 backbone.

Table 3 shows the results of the 5-way 1-shot and 5-way 5-shot classification on *mini*-ImageNet and *CUB* using this method. The experimental results of 5-way 5-shot show that the accuracy of our method on *CUB* is improved by 0.92% compared to the DeepBDC model [11]. Our method outperforms the current SOTA few-shot classification methods on the 5-shot settings. It is well known that "the fewer images selected in each category (such as 1-shot), the higher the requirement for model optimization capability". The experimental results of the model proposed in this paper under 5-way 1-shot are also improved by different degrees. Specifically, our method improves by 4.23%, 1.34%, 1.32% and 0.47% compared with Baseline++ [39], Neg-Margin [10], Meta-Baseline [19] and DeepBDC [11] on *CUB*, respectively, and shows superiority in the 1-shot classification task.

In summary, our model exhibits superior performance over other SOTA methods [8–11, 16, 19, 25, 39] in both the 1-shot and 5-shot settings, particularly in the 1-shot case. The reason for this phenomenon is that as the number of support sets increases, the available information for each category becomes richer, but compared with other few-shot learning methods, our model makes

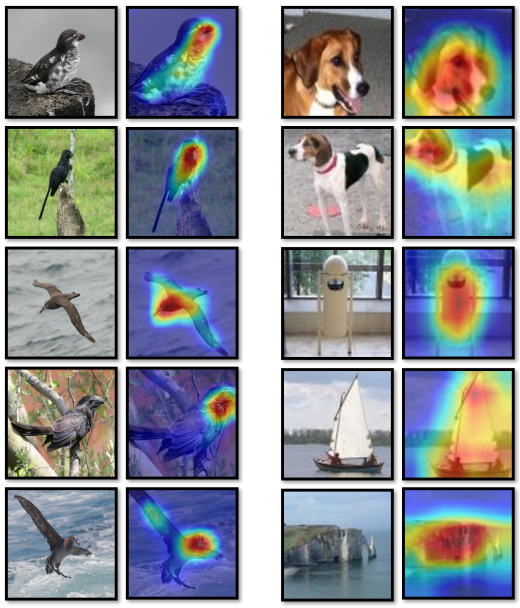


Fig. 6 Visualization of attention maps generated by MPHAM. (a) shows the focal region on the fine-grained dataset *CUB*, (b) indicates the focal region on *mini*-Imagenet.

more effective use of the information provided by the support sets. Through more noteworthy task-related information capture and discriminative and diversity enhancement, higher classification accuracy is ultimately obtained in the 1-shot setting.

4.5 A Domain-Specific Application: *FPV*

For domain-specific real-world tasks such as agricultural plant disease classification, large-scale sample collection is not feasible because experienced agricultural experts are often required to label them. There are some open source datasets available on the Internet, but their classification is generally detailed only to the disease category. The Fine-PlantVillage dataset (*FPV*) targeted in this paper classifies the category to the disease severity, which is significant for solving practical agricultural problems. Therefore, this section focuses on the experiments and discussion of *FPV* dataset. Figure 5 shows some examples of the dataset, with the original *PV* dataset examples on the left and the fine-grained dataset *FPV* on the right. *PV* is a plant disease image dataset with 54,305 images of plant diseases, including 38 categories of disease leaves from 13 species of plants. *FPV* has further refined the categories based on *PV*. *FPV* includes 61 classes with a total of 45,285 images. It can be seen from Figure 5 that the same picture has different category labels under different fine-grained requirements. We are faced with the challenge of fine-grained classification on *FPV* datasets. And it is extremely difficult to classify the severity of diseases within the same disease category.

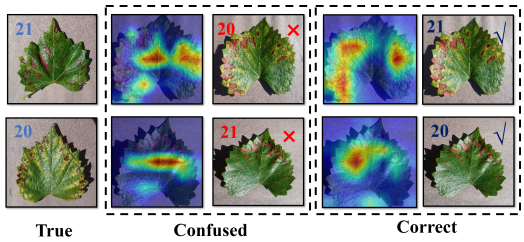


Fig. 7 Difficult category pairs to classify. The first column is the labeled training data. The red numbers in the middle column represent the incorrectly predicted categories, and the third column shows the correct prediction results.

For this dataset, 5-way experiments are conducted with both 1-shot and 5-shot trials. In the *FPV* dataset we randomly select 30 classes as base classes and the remaining classes as novel classes to perform the few-shot learning task. As shown in Table 3, it has obtained competitive performance in the 5-way 5-shot compared to many recent, more complex methods. Compared to methods with the same complexity, our work focuses on improving the performance of few-shot classification on fine-grained image datasets and shows significant advantages on *CUB* and *FPV*. The classification accuracy improved from 89.45% to 90.25% in the *FPV* 5-shot experiments and from 79.00% to 79.84% in 1-shot experiments.

In addition, the focus regions of some examples in *FPV* are visualized. As shown in Figure 8, we randomly select several categories of samples and visualize the heat map of feature weights. It can be noticed that our model displays different heat map focus distributions on each category, extracting different features. As you can see, our method can capture the key regions of the object, which helps the model extract discriminant features for classification.

Figure 9 shows the confusion matrix of our method and the baseline model on the *FPV*

Table 5 Comparison of the state-of-the-art few-shot classification algorithms on the *FPV* dataset. The best results are highlighted in bold.

Methods	*FPV*	
	5way1shot	5way5shot
Optimization-based		
MAML [16]	69.96 ± 0.46	82.84 ± 0.40
Metric-based		
Matching N [25]	77.93 ± 0.44	88.63 ± 0.31
Relation N [9]	74.00 ± 0.42	85.86 ± 0.36
Prototypical N [8]	74.22 ± 0.43	86.70 ± 0.30
Baseline [39]	71.96 ± 0.37	87.25 ± 0.22
Baseline++ [39]	76.11 ± 0.40	88.73 ± 0.31
Meta-Baseline [19]	78.25 ± 0.41	88.76 ± 0.26
Neg-Margin [10]	78.06 ± 0.46	88.48 ± 0.37
FEAT [40]	76.25 ± 0.41	88.02 ± 0.24
BML [41]	77.21 ± 0.63	89.33 ± 0.29
DeepEMD [42]	76.69 ± 0.47	87.92 ± 0.34
DeepBDC [11]	79.00 ± 0.52	89.45 ± 0.26
Ours	**79.84 ± 0.33**	**90.25 ± 0.23**

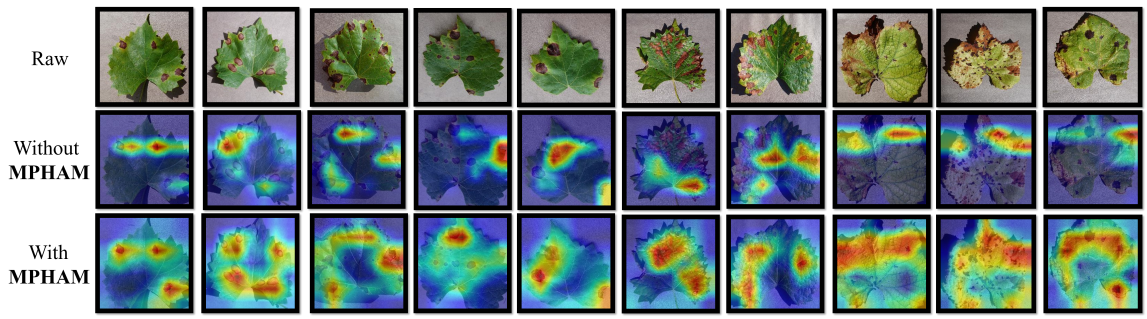
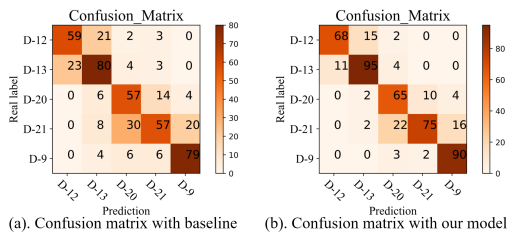


Fig. 8 Visualization of the focal region localization on the *FPV* dataset. The second row shows the focal region without MPHAM, and the third row shows the focal region with MPHAM.



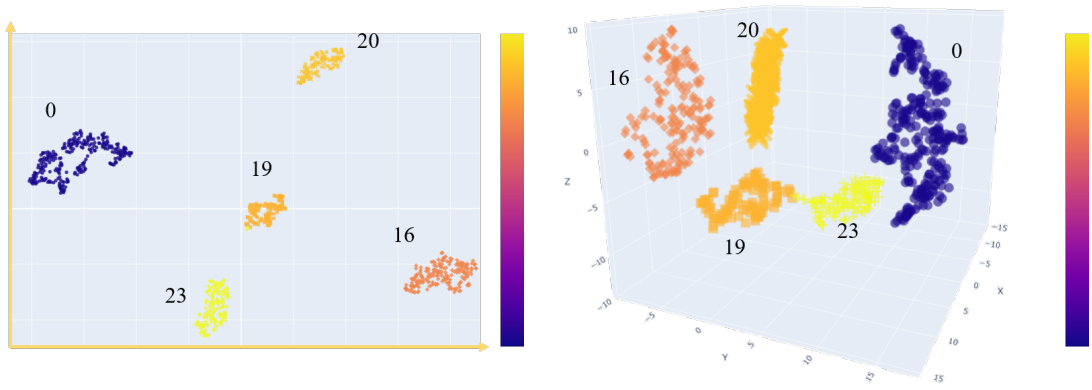
(a). Confusion matrix with baseline (b). Confusion matrix with our model

Fig. 9 Confusion matrix of the baseline and our model. D-12, D-13, D-20, D-21 and D-9 are five representative categories. Each column in the matrix represents the prediction result. Each row represents the real label.

dataset. It can be seen that the proposed method greatly improves the classification accuracy on class 20 and 21, and the possibility of class 13 being classified as class 12 is reduced. We perform t-SNE on the high-dimensional representations of easy-to-classify samples and hard-to-classify samples under our model. For the convenience of viewing, we only show the dimensionality reduction results for 5 classes of samples. Figure 10 shows the dimensionality reduction results for the easily classified samples, corresponding to the coarse-grained images in PV. The t-SNE results for the fine-grained images are presented in Figure 11, whose class pairs "class20-class21" and "class12-class13" have extremely similar feature representations.

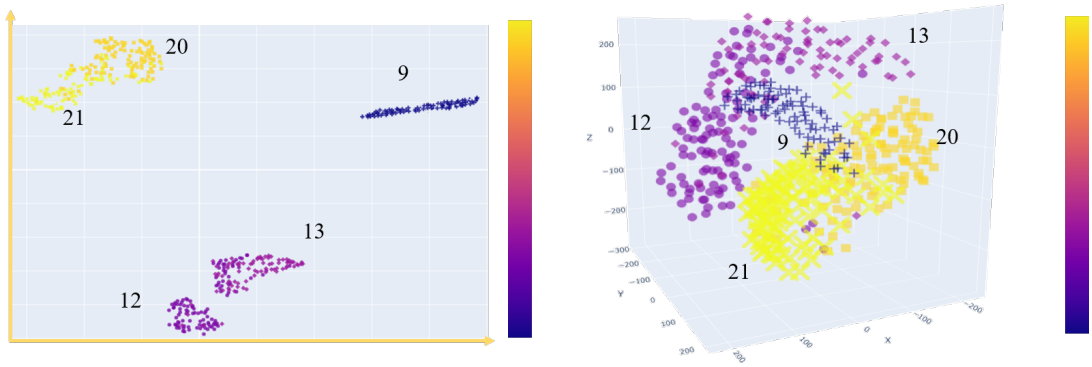
Figure 7 shows the difficult category pairs to distinguish. For example, samples of class 20 are generally classified as class 21. It can be seen from Figure 7 that the difference between these categories is very small. This similarity even confuses agricultural experts to distinguish them.

Summarizing the above experimental phenomena, there is also a fine-grained recognition problem for few-shot classification. Compared with



(a). 2D t-SNE visualization (b). 3D t-SNE visualization

Fig. 10 t-SNE visualization of 5 coarse-grained classes. (a) shows the 2D t-SNE visualization, (b) shows the 3D t-SNE visualization.



(a). 2D t-SNE visualization (b). 3D t-SNE visualization

Fig. 11 t-SNE visualization of 5 fine-grained classes. (a) shows the 2D t-SNE visualization, (b) shows the 3D t-SNE visualization.

the existing models, the model proposed in this paper achieves significant performance improvement, which also fully reflects the advantages of our method:

- With MPHAM and DDCM modules, the ability to capture key regions is obtained, and the discriminability and diversity of classification results are simultaneously improved. Through this method, the information utilization rate of few samples is more effectively improved. Our approach is of great value in the field of fine-grained few-shot classification.
- For domain-specific applications where data acquisition is difficult, such as on fine-grained plant disease datasets, our model has typical applicability, and the effect is significantly improved on *FPV*. The few-shot model proposed in paper has important significance in intelligent agricultural disease classification.

5 Conclusion

Originating from real-world needs, this paper focuses on the fine-grained few-shot plant disease classification problem by exploring attentional features of several labeled samples. In order to effectively capture the fine details of fine-grained plant diseases, this paper further proposes a multi-perspective hybrid attention module (MPHAM), which focuses on the global information of the image from different angles using instance-attention, region-attention, soft-attention, and patch-hard-attention. Aiming at improving the discriminability and diversity of the classifier, this paper introduces DDCM in the loss function to constrain the Batch Nuclear-norm of the classification matrix, which effectively improves the classification accuracy of the hard-to-classify samples. Extensive experiments are carried out to verify the effectiveness of the proposed

module. On a fine-grained plant disease dataset, this paper also completes the few-shot classification of natural images in practical industrial applications. Undoubtedly, the method presented in this paper is a valuable supplement to the fine-grained few-shot classification problem in the field of intelligence agricultural applications.

Declarations

- **Conflict of Interest.** The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.
- **Data Availability.** Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Acknowledgments. This work was supported by the National Key R&D Program of China (No. 2021ZD0110901).

References

[1] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[3] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

[4] Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)

[5] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[6] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

[7] Koch, G., Zemel, R., Salakhutdinov, R., *et al.*: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2, p. 0 (2015). Lille

[8] Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)

[9] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., 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)

[10] Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., Hu, H.: Negative margin matters: Understanding margin in few-shot classification. In: European Conference on Computer Vision, pp. 438–455 (2020). Springer

[11] Xie, J., Long, F., Lv, J., Wang, Q., 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)

[12] Ferentinos, K.P.: Deep learning models for plant disease detection and diagnosis. Computers and electronics in agriculture **145**, 311–318 (2018)

[13] Selvaraj, M.G., Vergara, A., Ruiz, H., Safari, N., Elayabalan, S., Ocimati, W., Blomme, G.: Ai-powered banana diseases and pest detection. Plant Methods **15**(1), 1–11 (2019)

[14] Aboneh, T., Rorissa, A., Srinivasagan, R., Gemechu, A.: Computer vision framework for wheat disease identification and classification using jetson gpu infrastructure. Technologies **9**(3), 47 (2021)

[15] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning, pp. 1842–1850 (2016). PMLR

[16] Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135 (2017). PMLR

[17] Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)

[18] Abbas, M., Xiao, Q., Chen, L., Chen, P.-Y., Chen, T.: Sharp-maml: Sharpness-aware model-agnostic meta learning. arXiv preprint arXiv:2206.03996 (2022)

[19] Chen, Y., Wang, X., Liu, Z., Xu, H., Darrell, T.: A new meta-baseline for few-shot learning (2020)

[20] Fu, J., Zheng, H., Mei, T.: 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, pp. 4438–4446 (2017)

[21] Sun, X., Xv, H., Dong, J., Zhou, H., Chen, C., Li, Q.: Few-shot learning for domain-specific fine-grained image classification. IEEE Transactions on Industrial Electronics **68**(4), 3588–3598 (2020)

[22] Wei, X.-S., Luo, J.-H., Wu, J., Zhou, Z.-H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing **26**(6), 2868–2881 (2017)

[23] Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217 (2017)

[24] Behera, A., Wharton, Z., Hewage, P.R., Bera, A.: Context-aware attentional pooling (cap) for fine-grained visual classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 929–937 (2021)

[25] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29** (2016)

[26] Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 751–766 (2018)

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

[28] Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

[29] Cui, S., Wang, S., Zhuo, J., Li, L., Huang, Q., Tian, Q.: Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3941–3950 (2020)

[30] Song, J., Shen, C., Yang, Y., Liu, Y., Song, M.: Transductive unbiased embedding for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1024–1033 (2018)

[31] Zhuo, J., Wang, S., Cui, S., Huang, Q.: Unsupervised open domain recognition by semantic discrepancy minimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 750–759 (2019)

[32] Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5982–5991 (2019)

[33] Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 289–305 (2018)

[34] Fazel, M.: Matrix rank minimization with applications. PhD thesis, PhD thesis, Stanford University (2002)

[35] Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review **52**(3), 471–501 (2010)

[36] Srebro, N., Rennie, J., Jaakkola, T.: Maximum-margin matrix factorization. Advances in neural information processing systems **17** (2004)

[37] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[38] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)

[39] Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C.F., Huang, J.-B.: A closer look at few-shot classification. arXiv preprint arXiv:1904.04232 (2019)

[40] Ye, H.-J., Hu, H., Zhan, D.-C., 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)

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

[42] Zhang, C., Cai, Y., Lin, G., 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 (2020)