electronics

MDPI

Article

Fine-Grained Few-Shot Image Classification Based on Feature Dual Reconstruction

Shudong Liu, Wenlong Zhong, Furong Guo, Jia Cong and Boyu Gu *

School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China; liushudong@tcu.edu.cn (S.L.); zhongwenlong2022@163.com (W.Z.); guofurong2022100@163.com (F.G.); congj2020@tcu.edu.cn (J.C.)

* Correspondence: guboyu@tcu.edu.cn

Abstract: Fine-grained few-shot image classification is a popular research area in deep learning. The main goal is to identify subcategories within a broader category using a limited number of samples. The challenge stems from the high intra-class variability and low inter-class variability of fine-grained images, which often hamper classification performance. To overcome this, we propose a fine-grained few-shot image classification algorithm based on bidirectional feature reconstruction. This algorithm introduces a Mixed Residual Attention Block (MRA Block), combining channel attention and window-based self-attention to capture local details in images. Additionally, the Dual Reconstruction Feature Fusion (DRFF) module is designed to enhance the model's adaptability to both inter-class and intra-class variations by integrating features of different scales across layers. Cosine similarity networks are employed for similarity measurement, enabling precise predictions. The experiments demonstrate that the proposed method achieves classification accuracies of 96.99%, 98.53%, and 89.78% on the CUB-200-2011, Stanford Cars, and Stanford Dogs datasets, respectively, confirming the method's efficacy in fine-grained classification tasks.

Keywords: image processing; few shot; feature fusion; attention mechanism; fine-grained image

Citation: Liu, S.; Zhong, W.; Guo, F.; Cong, J.; Gu, B. Fine-Grained Few-Shot Image Classification Based on Feature Dual Reconstruction. *Electronics* **2024**, *13*, 2751. https://doi.org/10.3390/electronics13142751

Academic Editor: Chiman Kwan

Received: 11 June 2024
Revised: 5 July 2024
Accepted: 10 July 2024
Published: 13 July 2024

1. Introduction

In recent years, considerable advances have been made in image tasks using deep learning, leveraging rich image representations obtained from abundant annotated data to enhance classification performance. However, not all problems have access to vast amounts of data for training [1]. For instance, scenarios such as endangered animal recognition, military weapon identification, and rare disease detection often suffer from limited sample sizes. Moreover, due to substantial intra-class differences in such scenarios, traditional deep learning methods exhibit notable shortcomings in recognition accuracy. Therefore, there is a need to address these fine-grained differences using few-shot learning approaches. Fine-grained image classification, a critical branch of image classification, requires not only distinguishing between basic categories but also finely dividing subcategories [2]. Thus, the challenge lies in learning more diverse and discriminative feature representations from fewer images for fine-grained image few-shot classification tasks [3].

To tackle these issues, it is crucial to delve into discriminative regional features that contribute to precise classification. For example, Zhu et al. [4] applied meta-learning methods based on multiple attention mechanisms to fine-grained recognition, enabling the model to adaptively focus on discriminative areas in images through attention mechanisms, thereby improving classification accuracy. Wertheimer et al. [5,6] framed the fine-grained classification problem as a feature reconstruction problem in latent space, reducing the influence of fine-grained image characteristics by reconstructing features from support sets and query sets in different ways. Although the FRN [5] (Feature Reconstruction Network) learns more detailed feature regions compared to prototype networks [7], experimental analysis revealed that when applied to the "birds" dataset [8], the model, while focusing

on more discriminative regions, still exhibited inconsistencies in semantic information between any two samples of the same category. This is due to the significant intra-class differences and small inter-class differences in fine-grained datasets, leading to instability in extracting consistent semantic information during feature extraction by the FRN [5] due to noise factors such as individual differences, shooting angles, and lighting conditions. Zhang et al. [9] proposed an innovative Earth Mover's Distance (EMD) distance measurement method, which evaluates the similarity between the support set and query set images by computing the optimal matching cost between image blocks. Li et al. [10] further proposed an algorithm called BSNet (Bi-Similarity Network). BSNet constructs more accurate feature maps by learning and integrating two similarity measures based on different features, enabling the model to extract more discriminative feature representations for fine-grained few-shot tasks, effectively reducing similarity bias and significantly enhancing the model's generalization ability. EGNN [11] is a few-shot learning method based on graph convolutional networks. It leverages intra-class similarity and inter-class differences to iteratively update edge labels, which are then used to predict node labels. The model simultaneously trains the features of both the support set and the query set using a transductive method, thereby optimizing the performance of few-shot classification. The advantage of this approach is that it is suitable for transfer between different categories, and after transfer, it does not require the retraining of parameters. Tang et al. [12] proposed the PMRC framework, which uses a deep navigator to generate discriminative regions from images and then constructs a graph using these regions. The graph is aggregated through message passing to obtain the classification results. Chang et al. [13] introduced an erudite FGVC model jointly trained with multiple different datasets to reduce negative transfer between datasets and promote positive transfer. Lyu et al. [14] presented a Siamese transformer with hierarchical concept embedding (STrHCE), which is made up of two transformer subnetworks that share all settings and are provided with hierarchical semantic information at multiple concept levels for fine-grained picture embeddings. Wu et al. [15] introduced a feature self-reconstruction mechanism to help the model explore more discriminative features. However, due to its high dependency on the quality of the initial features [16], the model cannot fully represent the underlying features. This issue remains a major challenge in fine-grained classification.

Fine-grained few-shot classification algorithms mainly fall into two categories: meta-learning-based methods and metric-learning-based methods [17]. Meta-learning-based methods mainly utilize existing knowledge and experience to enable the model to quickly adapt and handle new classification tasks. For example, Finn et al. [18] proposed the Model-Agnostic Meta-Learning (MAML) method, which aims to find parameters in the network that are sensitive to each task. By fine-tuning these parameters, the model can achieve rapid convergence. Rusu et al. [19] proposed the LEO few-shot learning algorithm, which introduces a low-dimensional latent space and performs inner-loop parameter updates within this latent space. In contrast, metric-learning-based methods focus more on optimizing feature embeddings to improve the spatial distribution of feature vectors [20]. Therefore, metric learning methods emphasize the relative relationships between samples rather than absolute feature representations, which perform well in image classification tasks by helping the model better understand image content, extract discriminative features, and enable the model to obtain good parameter distributions with only a small number of samples [21]. In fine-grained few-shot classification tasks, due to the subtle differences between categories, accurately measuring the similarity between samples becomes a crucial issue. However, many current metric-based fine-grained image classification methods rely solely on a single similarity measurement approach. In situations with relatively limited samples, this singularity may lead to bias in similarity calculations. Moreover, convolutional neural networks often struggle to deeply explore more discriminative fine-grained features due to their limited capacity to capture sufficient local details and fuse cross-scale features [15]. To address these challenges, we present a fine-grained few-shot image classification approach based on feature dual reconstruction. This approach uses

metric learning to address the fine-grained few-shot picture categorization problem. The primary contributions of this study include:

(1) We proposed an MRA Block and designed a Mixed Attention Block (MAB) and an Overlapping Cross-Attention Block (OCAB). The MRA Block consists of MAB and OCAB in series, integrating channel attention and window-based self-attention mechanisms to enhance cross-window connections and information aggregation capabilities. Combined with multiple residual mechanisms, this design merges features from the main and side branches, enabling the model to comprehensively understand data distribution and accurately respond to feature changes in different local regions.

(2) We proposed a Dual Reconstruction Feature Fusion (DRFF) module, which applies feature fusion to feature reconstruction through an alternating integration approach. This module uses the support set to reconstruct the query set to increase inter-class variation, while using the query set to reconstruct the support set to reduce intra-class variation. Additionally, we introduced the Transformer encoder and attention feature fusion module (AFFM) to enable the model to dynamically adapt to inter-class and intra-class variations, enhancing its ability to capture key information at different scales.

(3) We performed in-depth ablation investigations and evaluations on three popular fine-grained categorization datasets. These careful evaluations proved our suggested method's improved performance and robustness, underscoring its efficacy in tackling the difficulties associated with fine-grained few-shot categorization.

The manuscript is structured as follows: Section 1 presents an overview of the existing literature as well as this paper's primary contributions. Section 2 contains a full discussion of the proposed model structure. Section 3 outlines the datasets utilized for model training and evaluation, as well as an in-depth experimental examination of three standard datasets. Section 4 examines current experimental findings and prospective research directions. Section 5 presents the general conclusion.

2. Methods

2.1. Algorithm Structure

This study offers a fine-grained image classification method based on feature dual reconstruction, with the goal of properly measuring the similarity of feature maps and learning more discriminative feature representations in response to the features of fine-grained few-shot classification problems. The specific structure is illustrated in Figure 1.

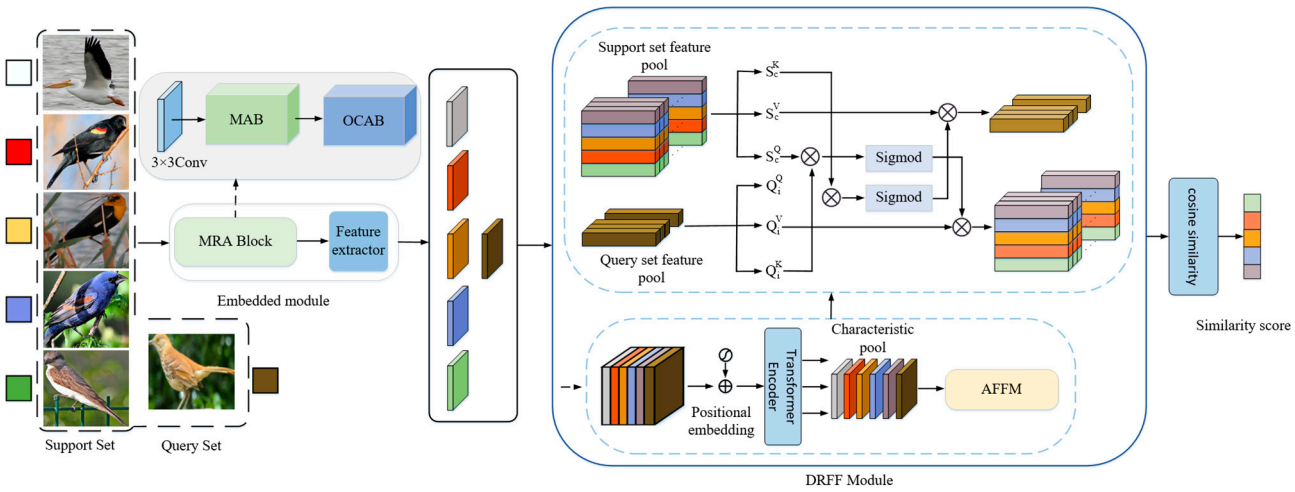


Figure 1. Fine-grained small-sample image classification model based on feature dual reconstruction.

Our model consists of four major modules. The first is the embedding module, which extracts deep convolutional picture features. It has an MRA Block and a feature extractor,

which might be a standard convolutional neural network or a residual network. The second module is the Dual Reconstruction Feature Fusion (DRFF), which uses a self-attention mechanism to reconstruct each image's convolutional features before refining them with the AFFM to produce more expressive features. The third module is the mutual feature reconstruction module, which reconstructs sample features bidirectionally. This module reconstructs both the support sample and the query sample.

Compared to the current unidirectional reconstruction, which only enhances inter-class feature variations, the bidirectional reconstruction introduces an additional capability—reducing intra-class feature variations. The fourth module is the cosine similarity measurement module, which calculates the distance between the original and reconstructed sample features. The classification of query samples is then based on the weighted sum of these two distances.

For a new task, support set images and query set images are first input into the embedding module for image feature extraction. Within the embedding module, a 3×3 convolutional layer, a Mixed Attention Block (MAB), and an Overlapping Cross-Attention Block (OCAB) are combined to more effectively extract features from input images. Subsequently, through the DRFF module, while reshaping and reconstructing features, the model's perception of multi-scale features is enhanced. Finally, the cosine similarity of the reconstructed features of the support set and query set is calculated to obtain the final similarity score.

2.2. Embedding Module

The embedding module in this paper consists of an MRA Block and a feature extractor. The MRA Block was built of a 3×3 convolutional layer, a Mixed Attention Block (MAB), and an OCAB. The input channel number of the MAB was set to 180, the OCAB had 6 attention heads, and the window size was set to 16.

2.2.1. MRA Block

The MRA Block enhances the representation capability of image features through fine feature extraction and the integration of attention mechanisms. The internal structure of the MRA Block consists of three main components: a 3×3 convolution layer for initial feature extraction, followed by a Mixed Attention Block (MAB), which combines channel attention and multi-head self-attention, and an Overlapping Cross-Attention Block (OCAB), composed of two parallel overlapping cross-attention layers (OCAs) to facilitate cross-region information exchange.

In the module, channel attention is used to identify key channels in the image and assign them higher weights, thereby enhancing the model's response to significant features in the image. The structure of the channel attention is shown in Figure 2.

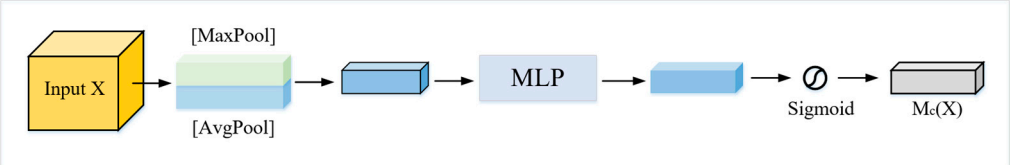


Figure 2. Channel attention structure.

Multi-head self-attention allows the model to capture input from multiple subspaces at the same time. Each head learns an attention distribution that highlights the significance of different features, enriching the model's feature representations. The architecture of multi-head self-attention is depicted in Figure 3.

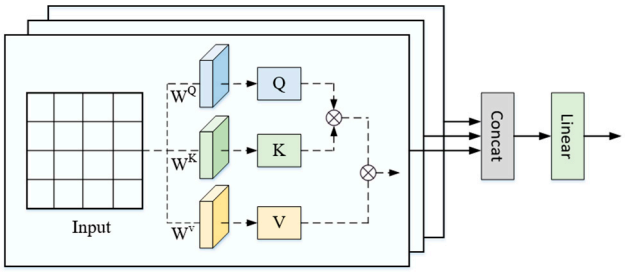


Figure 3. Multi-head self-attention structure.

2.2.2. Mixed Attention Block

We added a convolution block based on channel attention into the normal Transformer block to improve the network's representation capability, since convolution can help Transformers obtain better visual representations [22–26]. The Window-based Multi-Head Self-Attention (W-MSA) module and the Channel Attention block were serially inserted into the basic Swin Transformer block, as seen in Figure 4. The complete computing method of the Mixed Attention Block for a given input feature X is as follows in order to avoid conflicts in visual representations between the Multi-Head Self-Attention module and the Channel Attention block [27,28].

$$X_N = \mathrm{LN}(X) \tag{1}$$

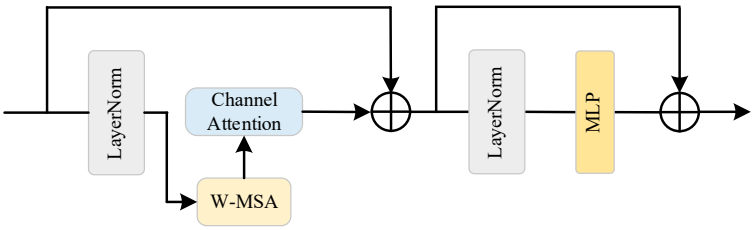


Figure 4. MAB structure diagram.

X_N represents the result of the input feature X after LayerNorm. Next, X_N is processed by the W-MSA module, which divides the features into local windows and independently applies the multi-head self-attention mechanism within each window to capture local dependencies.

$$X_M = \text{W-MSA}(X_N) + \mathrm{CAB}(\text{W-MSA}(X_N)) \tag{2}$$

In this step, X_M is the output of the MAB. W-MSA(X_N) represents the outcome of the W-MSA, and CAB is the Channel Attention Block that processes the W-MSA(X_N) output to highlight key channel features. The final feature representation of the Mixed Attention Block is obtained by adding the output of the Channel Attention to the output of the W-MSA.

$$Y = \mathrm{MLP}(\mathrm{LN}(X_M)) + X_M \tag{3}$$

X_M is first normalized through a LayerNorm operation, then fed into a Multi-Layer Perceptron (MLP). The MLP refines the features with two fully connected layers and a GELU activation function. The output of the MLP is then combined with the original X_M through a residual connection, preserving the information flow and improving the model's capability to learn complex features.

Channel attention assigns importance weights to different channels of the feature map. First, global average pooling (GAP) and global max pooling (GMP) are applied to the input feature X, resulting in two different representations that capture global and local features, respectively. These representations are then transformed by learnable weights W_1 and

W_2, and normalized through the sigmoid function σ, generating the attention scores. The formula for calculating the attention scores is:

$$\text{Attention} = \sigma(W_1 GAP(X) + W_2 GMP(X)) \tag{4}$$

The Channel Attention Block parameters are as follows: a 3×3 convolution kernel, a stride of 1, and a number of output channels equal to the input channels. For window-based self-attention, the window size is set to 7×7, with 6 attention heads and a linear layer dimension of 512.

The window-based self-attention mechanism not only considers the interaction of information within local windows but also achieves adaptive weighting of different positional information through the attention mechanism, further enhancing the model's representation capabilities. For input features of the size $H \times W \times C$, they are divided into multiple local windows. This operation enhances computational efficiency by limiting the self-attention mechanism's scope to local windows, reducing its computational complexity. Self-attention is then calculated within each local window. Linear mapping through fully connected layers is performed on the features of each window to transform the input features from their original space to a new space, producing the query matrix Q, key matrix K, and value matrix V. Following this, the standard self-attention calculation is executed: first, the dot product between Q and K is calculated, followed by normalization with Softmax to obtain the attention weight matrix. This weight matrix is then used to perform weighted aggregation with V, yielding new feature representations for each position. The window-based self-attention calculation process is as follows:

$$\text{Attention}(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}}) \tag{5}$$

In this calculation, d stands for the dimension of the query key. The shifted window method proposed by Liu, Lin, and Cao [29] is employed. Figure 4 illustrates the structure of the MAB.

2.2.3. Overlapping Cross-Attention Block

To better establish inter-window connections for aggregating inter-window information, this paper designs an Overlapping Cross-Attention Block to enhance feature representation by directly establishing inter-window connections. Similar to the standard Swin Transformer block [30], it consists of two parallel Overlapping Cross-Attention layers (OCAs) and a multi-layer perceptron.

Each OCA layer processes as follows: for each window of the size $M \times M$, the query (Q), key (K), and value (V) matrices are computed; then, the attention weights are obtained by calculating the dot product of Q and K, and the weighted feature representation is generated by multiplying these weights with the V matrix; next, the multi-head self-attention mechanism aggregates features within each window, utilizing the overlapping window design to achieve cross-region feature integration. This design helps the model maintain local details while understanding the global contextual information. The overlapping window partition design is shown in Figure 5. Finally, an MLP consisting of two fully connected layers and a GELU activation function is applied to further process the aggregated features, with layer normalization performed before the attention layer and the MLP, and dropout was used after each fully connected layer in the MLP to prevent overfitting.

OCA calculates the keys/values from a larger field, using different window sizes to partition the input features, as shown in Figure 5. Specifically, for the inputs X_Q, X_K, and X_V, X_Q is divided into non-overlapping windows of the size $M \times M$, while X_K and X_V are expanded into overlapping windows of the size $M_0 \times M_0$, where "expanded into windows" means we adopt a strategy to ensure a certain overlap region between adjacent windows, which is illustrated with red boxes in Figure 5. As M increases, M_0 will overlap. This overlapping design allows information to flow at the window boundaries, thereby

achieving inter-window information aggregation. The calculation formula for M_0 is as follows.

$$M_0 = (1 + \gamma) \times M \tag{6}$$

Here, M is the sliding window size of the standard window partition, M_0 is the sliding window size of the overlapping window, and γ is a constant that controls the overlap size. This design facilitates the seamless integration of information across different windows, thereby enhancing the model's ability to capture global context. Moreover, within each overlapping window, the attention mechanism calculates attention scores and applies them to the values, thereby enhancing relevant features and suppressing less important ones. This configuration allows the model to efficiently gather and disseminate information throughout the feature map, yielding a more robust and detailed feature representation. Figure 6 depicts the structure of the overlapping Cross-Attention Block.

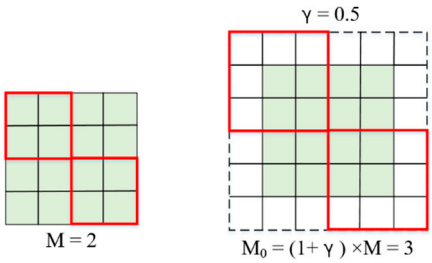


Figure 5. Overlapping window partitions of OCA.

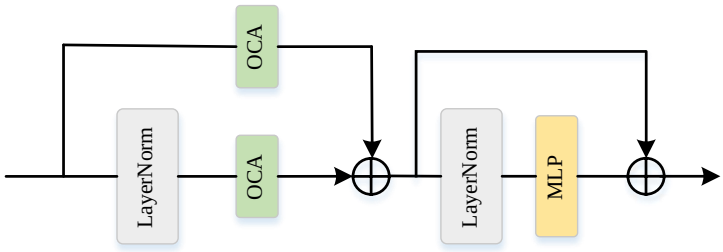


Figure 6. OCAB structure diagram.

2.3. Dual-Reconstruction Feature Fusion Module

The Dual-Reconstruction Feature Fusion module (DRFF) achieves deep feature fusion and reconstruction by integrating the Transformer encoder, attention feature fusion module (AFFM), and mutual reconstruction module. This module's design aims to enhance the model's ability to capture critical information in fine-grained image classification tasks.

The DRFF module first receives specially processed input through a Transformer encoder. Unlike traditional vision transformers, we do not directly input the image patch sequences. Instead, we compute the sum of local feature sequences and their corresponding spatial position embeddings, $z_i = [x_{1i}, x_{2i}, \ldots, x_{ri}] + E_{\text{pos}}$. This approach allows the model to incorporate spatial position information while considering the features of each position.

Subsequently, the output of the Transformer encoder is passed to the AFFM, which performs feature fusion based on the standard self-attention operation. Existing feature fusion methods, such as those in SKNet [31] and ResNeSt [32], typically use simple initial integration when merging features, which fails to achieve true cross-layer feature fusion. To address this issue, we designed an alternating integration method. By combining initial integration with another attention module, we obtain more precise feature representations and introduce this alternating integration approach during the feature reconstruction stage. The distinctive aspect of the AFFM is its internal structure, which includes two Multi-Scale Channel Attention Modules (MS-CAMs). These modules capture channel dependencies at different scales, providing the model with richer and more detailed feature representations. Finally, the DRFF module reconstructs the features of the support set and query set through

a mutual reconstruction strategy. This process not only enhances the model's robustness to intra-class variations but also improves its sensitivity to inter-class differences. The overall structural design of the DRFF module is illustrated in Figure 1.

2.3.1. Multi-Scale Channel Attention Module

The MS-CAM enhances the representation capability of feature maps by capturing channel dependencies at multiple scales [33]. To maintain the lightweight nature of the MS-CAM during feature fusion, the global context information and local context information within the attention module are fused. Since pointwise convolution can interact with channels at each spatial location, we chose pointwise convolution as a tool for aggregating local channel context, and used ReLU after each convolution layer to introduce non-linearity. Batch Normalization (BN) was applied after each convolution layer to stabilize training. The BN settings were as follows: momentum: 0.9, epsilon: 1×10^{-5}. The internal structure of the MS-CAM is detailed in Figure 7.

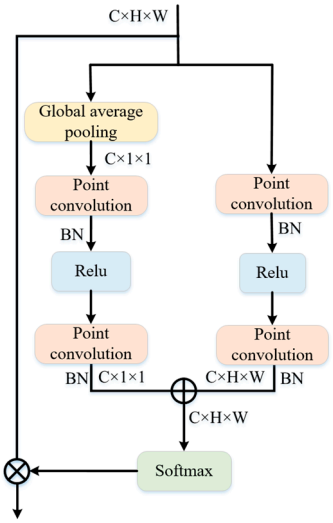


Figure 7. MS-CAM structure diagram.

2.3.2. Attention Feature Fusion Module

The AFFM shows good application in many typical cases, such as feature fusion inside Inception layers and generated by long and short skip connection layers. Conventional attention feature fusion techniques frequently involve merely adding the original features element by element before using the fused features as the attention module's input. However, the ultimate distribution of fusion weights may be directly impacted by such processing.

In contrast to traditional methods, our designed attention feature fusion module comprises two MS-CAMs within its internal structure. This design enables our module to more efficiently capture key information at different scales and generate more expressive features through weighted fusion. In greater detail, each MS-CAM in our module first processes the input features by capturing multi-scale channel dependencies. The global context information is obtained using global average pooling, while the local context information is aggregated using pointwise convolutions. ReLU activations are applied after each convolution layer to introduce non-linearity, allowing the model to learn more complex feature representations while maintaining high computational efficiency, thus aiding in faster convergence. In the MS-CAM, ReLU is used after the pointwise convolution layer to ensure that important feature variations can be effectively captured and amplified while aggregating local contextual information. Subsequently, Batch Normalization (BN) is applied to stabilize the training process. The BN settings are as follows: momentum: 0.9, epsilon: 1×10^{-5}.

After processing the features through the two MS-CAMs, the resulting feature maps were combined using an adaptive weighting mechanism. This mechanism assigns weights to each scale's features based on their importance, allowing for a more nuanced and effective fusion of the multi-scale information. The internal structure of the attention feature fusion module is detailed in Figure 8.

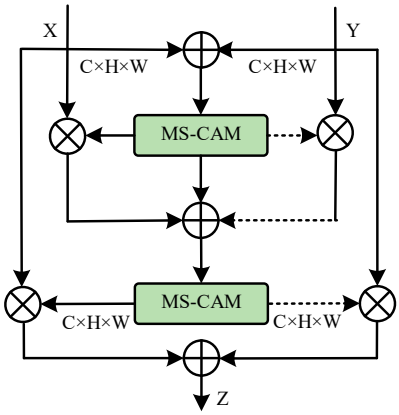


Figure 8. Structure diagram of attention feature fusion module.

3. Experiments

3.1. Dataset and Experimental Setup

3.1.1. Dataset

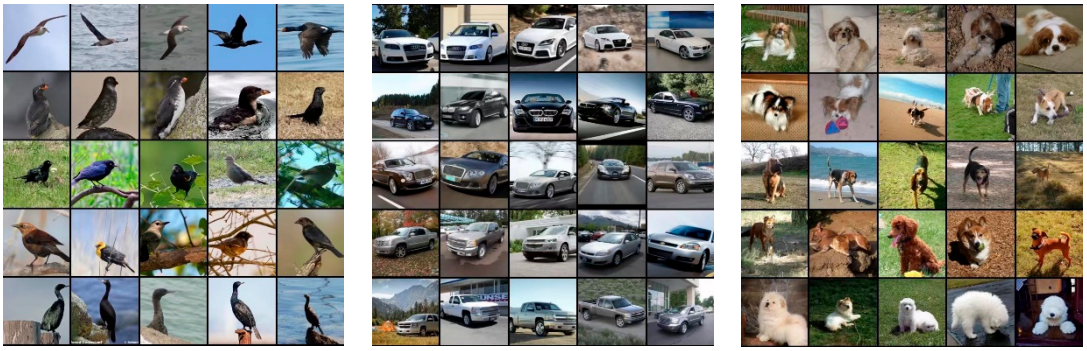
The experiments in this study utilized the CUB-200-2011, Stanford Cars (Cars), and Stanford Dogs (Dogs) datasets. These datasets are well known benchmarks in fine-grained image classification, characterized by significant intra-class variability and minimal inter-class differences. They offer a wide range of categories and samples, making them ideal for assessing the proposed method's effectiveness. Before conducting the experiments, all image samples were resized to 224 × 224 pixels to standardize the processing and fit our model input requirements. This resizing process maintained the original aspect ratio and content of the images using interpolation methods to minimize quality loss. We chose this size because it is common in image classification tasks and allows the model to capture sufficient details while controlling computational complexity. Additionally, we strictly divided the training set, validation set, and test set in a 2:1:1 ratio, ensuring that the categories were non-overlapping among these sets to guarantee the accuracy of the experiments.

CUB-200-2011 is a classic fine-grained image classification dataset consisting of 200 categories with a total of 11,788 bird images. The challenge of this dataset lies in the subtle differences between different bird species and the significant appearance variations within the same species.

Stanford Cars (Cars) contains 16,185 images of 196 types of cars. The dataset is characterized by a wide variety of car types, with significant appearance differences between cars of the same type.

The Stanford Dogs (Dogs) dataset contains 20,580 photographs of 120 different dog breeds from around the world. This dataset is designed specifically for fine-grained image classification. It was initially created to address the complex task of distinguishing between similar dog breeds. This task is particularly challenging because some breeds have very subtle differences, mainly in coloration and maturity.

To further enhance the clarity of our description, we provide a representative set of images from the three datasets below. Each dataset is represented by 25 images from different categories, visually illustrating the unique challenges and characteristics of these datasets. Figure 9a shows bird images from the CUB-200-2011 dataset; Figure 9b shows car images from the Stanford Cars dataset; and Figure 9c shows dog breed images from the Stanford Dogs dataset.



(**a**) (**b**) (**c**)

Figure 9. Few-shot classification datasets. (**a**) Sample image from the CUB-200-2011 dataset; (**b**) Sample image from the Stanford Cars dataset; (**c**) sample image from the Stanford Dogs dataset.

3.1.2. Experimental Setup

The experiments were carried out on a system with an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz, 32 GB RAM, and an Nvidia RTX3090 24 GB GPU, using the PyTorch 12.1.0 framework on Ubuntu. We performed classification tasks on three public datasets. The models were trained with the SGD optimizer, using Nesterov momentum (set to 0.9), a weight decay of 5×10^{-4}, and over 1200 epochs with an initial learning rate of 0.1. To enhance training stability, common data augmentation techniques, including center cropping, random horizontal flipping, and color jittering, were applied.

The best-performing models were chosen based on validation set results, which were assessed every 20 epochs. During testing, 10,000 tasks were randomly sampled to thoroughly evaluate the proposed method. The final test result was determined by calculating the average accuracy within a 95% confidence interval.

To evaluate the inference efficiency of our model, we measured the processing time per image on the three benchmark datasets. For a fair evaluation, we measured the inference speed using a single CPU core and GPU without parallel processing. On the CPU, our model's average inference time was 334 ms per image, while on the GPU, the average inference time significantly reduced to 34.3 ms per image. As shown in Table 1, the method proposed in this study improves the model's classification performance with only a small increase in inference time.

Table 1. The average inference time on the three datasets.

Model	Inference Time	Model	Inference Time
Conv4	17.6 ms	ResNet-12	22.6 ms
Conv4+MRA	24.8 ms	ResNet-12+MRA	31.5 ms
Conv4+DRFF	21.7 ms	ResNet-12+DRFF	27.1 ms
Conv4+MRA+DRFF	26.9 ms	ResNet-12+MRA+DRFF	34.3 ms

3.2. Evaluation Metrics

This paper addresses the C-way K-shot problem in fine-grained few-shot image classification. In this scenario, images are first trained on a support set and then tested on a query set, with no overlap in categories between the two sets. Given that 1-shot learning represents an extreme case of few-shot learning, and 5-shot provides more information while still being limited, we used 5-way 1-shot and 5-way 5-shot as the primary evaluation metrics. For constructing the training, validation, and testing sets, each category in the query set consistently had 15 images. Specifically, in a 5-way 1-shot task, we randomly selected 5 images from the training set for the support set and prepared 75 images for the query set. During training, C different categories were randomly selected from the support set, with K samples extracted from each category. To thoroughly assess model

performance, we used mean accuracy (meanAcc) as the main evaluation metric, calculated as shown in Formula (7), where k represents the batch size during training and the number of classifications during testing.

$$\text{MeanAcc} = \left[\frac{1}{k}\sum_{j=1}^{k}\left[\frac{\sum_{i=1}^{C} T_i}{\sum_{i=1}^{C}(T_i + F_i)}\right]\right] \times 100\% \quad (7)$$

3.3. Optimization and Convergence Analysis

To observe the optimization process and accurately assess the convergence of the objective function, we designed a series of additional experiments. These experiments focused on evaluating the accuracy of 5-way 1-shot and 5-way 5-shot classification tasks across different datasets. The experimental results showed that the accuracy of the model in both 5-way 1-shot and 5-way 5-shot tasks gradually increased with the number of training iterations. This indicates that as training progresses, the model continually optimizes its feature learning and representation, thereby improving its ability to recognize new categories.

Additionally, we observed that the rate of accuracy improvement was faster during the initial stages of training. This is because the model easily learned simple patterns in the early phase. As the number of training iterations increased, the rate of improvement slowed down as the model began learning more complex patterns. However, even in the later stages of training, accuracy continued to improve steadily, demonstrating that our model has good learning capability and convergence.

Figure 10 displays the 5-way 1-shot classification accuracy achieved on the three datasets. The x-axis represents the training epochs, and the y-axis indicates the accuracy percentage. As shown in the figure, the accuracy demonstrates a steady increase with each epoch, reflecting the effective optimization of our model. Figure 11 presents the 5-way 5-shot classification accuracy, which also shows a clear upward trend over the course of training. Similar to the 5-way 1-shot scenario, the model exhibits an enhanced ability to learn and generalize as training progresses. The stabilization of the curves towards the end of the training period suggests that the model has converged to an optimal solution.

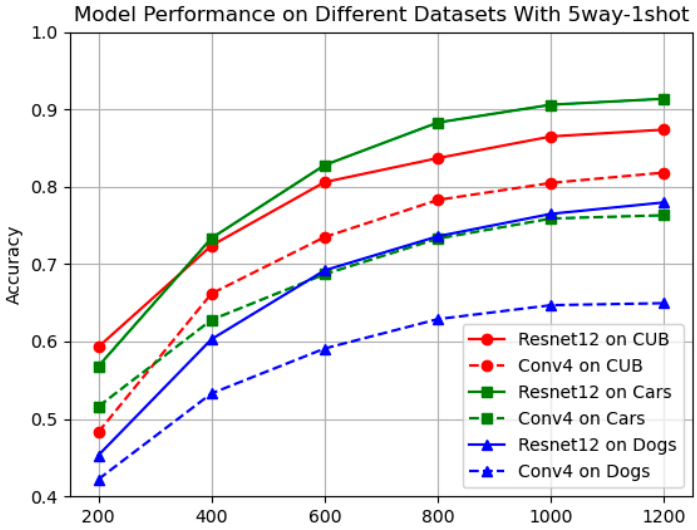


Figure 10. The 5-way 1-shot classification results on three different datasets.

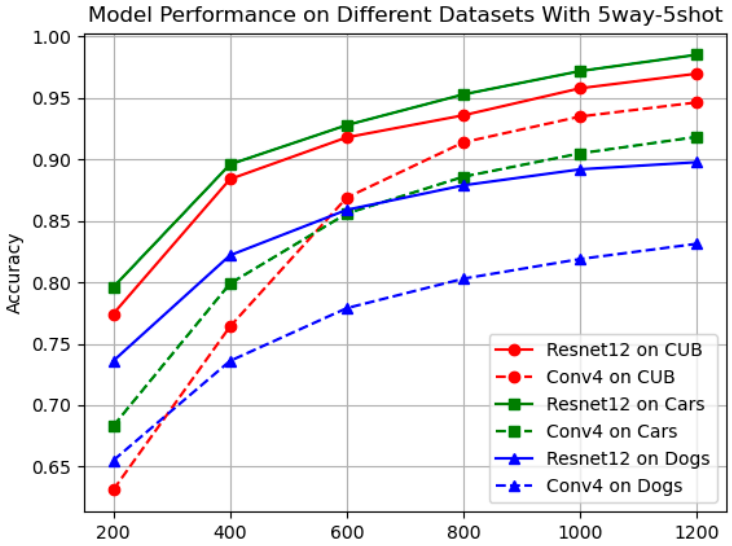


Figure 11. The 5-way 5-shot classification results on three different datasets.

3.4. Experimental Results

3.4.1. Comparison to the State of the Art

To demonstrate the effectiveness of our proposed approach, we quantitatively compared it to advanced fine-grained few-shot image categorization algorithms. Tables 2–4 show the classification results of our algorithm and other algorithms on the three datasets for the 5-way 1-shot and 5-way 5-shot tasks.

Table 2. Classification accuracy (%) on CUB-200-2011.

Method	Backbone	CUB-200-2011	
		5way-1shot	5way-5shot
ProtoNet [7]	Conv4	64.82 ± 0.23	85.74 ± 0.14
Relation [20]	Conv4	63.94 ± 0.92	77.87 ± 0.64
PARN [34]	Conv4	74.43 ± 0.95	83.11 ± 0.67
SAML [35]	Conv4	65.35 ± 0.65	78.47 ± 0.41
DeepEMD [9]	Conv4	64.08 ± 0.50	80.55 ± 0.71
LRPABN [36]	Conv4	63.63 ± 0.77	76.06 ± 0.58
BSNet [10]	Conv4	62.84 ± 0.95	85.39 ± 0.56
CTX [6]	Conv4	72.61 ± 0.21	86.23 ± 0.14
FRN [5]	Conv4	74.90 ± 0.21	89.39 ± 0.12
FRN+TDM [37]	Conv4	72.01 ± 0.22	89.05 ± 0.12
BiFRN [15]	Conv4	79.15 ± 0.19	92.21 ± 0.10
FS-FGIC [38]	Conv4	78.66 ± 0.46	89.43 ± 0.28
Ours	Conv4	81.83 ± 0.19	94.65 ± 0.10
ProtoNet [7]	ResNet-12	81.02 ± 0.20	91.93 ± 0.11
DeepEMD [9]	ResNet-12	75.59 ± 0.30	88.23 ± 0.18
CTX [6]	ResNet-12	80.39 ± 0.20	91.01 ± 0.11
FRN [5]	ResNet-12	84.30 ± 0.18	93.34 ± 0.10
FRN+TDM [37]	ResNet-12	85.15 ± 0.18	93.99 ± 0.09
BiFRN [15]	ResNet-12	85.44 ± 0.18	94.73 ± 0.09
FS-FGIC [38]	ResNet-12	86.14 ± 0.18	95.08 ± 0.09
Ours	ResNet-12	87.38 ± 0.18	96.99 ± 0.09

Table 3. Classification accuracy (%) on Stanford Cars.

Method	Backbone	Stanford Cars	
		5way-1shot	5way-5shot
ProtoNet [7]	Conv4	50.88 ± 0.23	74.89 ± 0.18
Relation [20]	Conv4	46.04 ± 0.91	68.52 ± 0.78
PARN [34]	Conv4	66.01 ± 0.94	73.74 ± 0.70
SAML [35]	Conv4	61.07 ± 0.47	88.73 ± 0.49
DeepEMD [9]	Conv4	61.63 ± 0.27	72.95 ± 0.38
LRPABN [36]	Conv4	60.28 ± 0.76	73.29 ± 0.58
BSNet [10]	Conv4	40.89 ± 0.77	86.88 ± 0.50
CTX [6]	Conv4	66.35 ± 0.21	82.25 ± 0.14
FRN [5]	Conv4	67.48 ± 0.22	87.97 ± 0.11
FRN+TDM [37]	Conv4	65.67 ± 0.22	86.44 ± 0.12
BiFRN [15]	Conv4	75.74 ± 0.20	91.58 ± 0.09
FS-FGIC [38]	Conv4	81.29 ± 0.45	91.08 ± 0.26
Ours	Conv4	76.32 ± 0.20	91.84 ± 0.09
ProtoNet [7]	ResNet-12	85.46 ± 0.19	95.08 ± 0.08
DeepEMD [9]	ResNet-12	80.62 ± 0.26	92.63 ± 0.13
CTX [6]	ResNet-12	85.03 ± 0.19	92.63 ± 0.11
FRN [5]	ResNet-12	88.01 ± 0.17	95.75 ± 0.07
FRN+TDM [37]	ResNet-12	88.92 ± 0.16	96.88 ± 0.06
BiFRN [15]	ResNet-12	90.44 ± 0.15	97.49 ± 0.05
FS-FGIC [38]	ResNet-12	88.96 ± 0.37	95.16 ± 0.20
Ours	ResNet-12	91.37 ± 0.18	98.53 ± 0.08

Table 4. Classification accuracy (%) on Stanford Dogs.

Method	Backbone	Stanford Dogs	
		5way-1shot	5way-5shot
ProtoNet [7]	Conv4	46.66 ± 0.21	70.77 ± 0.16
Relation [20]	Conv4	47.35 ± 0.88	66.20 ± 0.74
PARN [34]	Conv4	55.86 ± 0.97	68.06 ± 0.72
SAML [35]	Conv4	45.46 ± 0.36	59.65 ± 0.51
DeepEMD [9]	Conv4	46.73 ± 0.49	65.74 ± 0.63
LRPABN [36]	Conv4	45.72 ± 0.75	60.94 ± 0.66
BSNet [10]	Conv4	43.42 ± 0.86	71.90 ± 0.68
CTX [6]	Conv4	57.86 ± 0.21	73.59 ± 0.16
FRN [5]	Conv4	60.41 ± 0.21	79.26 ± 0.15
FRN+TDM [37]	Conv4	51.57 ± 0.23	75.25 ± 0.16
BiFRN [15]	Conv4	64.74 ± 0.22	81.29 ± 0.14
FS-FGIC [38]	Conv4	66.42 ± 0.50	81.23 ± 0.34
Ours	Conv4	64.96 ± 0.20	83.14 ± 0.09
ProtoNet [7]	ResNet-12	73.81 ± 0.21	87.39 ± 0.12
DeepEMD [9]	ResNet-12	70.38 ± 0.30	85.24 ± 0.18
CTX [36]	ResNet-12	73.22 ± 0.22	85.90 ± 0.13
FRN [5]	ResNet-12	76.76 ± 0.21	88.74 ± 0.12
FRN+TDM [38]	ResNet-12	78.02 ± 0.20	89.85 ± 0.11
BiFRN [15]	ResNet-12	76.89 ± 0.21	88.27 ± 0.12
FS-FGIC [38]	ResNet-12	75.50 ± 0.49	87.65 ± 0.28
Ours	ResNet-12	77.98 ± 0.18	89.78 ± 0.08

Experiments show that traditional methods such as Matching Networks and Prototypical Networks have certain limitations in fine-grained feature extraction, often leading to image misclassification. For example, in the CUB-200-2011 dataset, the albatross and the white-necked duck have high similarity in appearance and flight posture, which frequently results in misclassification. As shown in Figure 12, due to subtle differences in beak shape, feather patterns, and body structure, as well as intra-species variation, ecological niche similarity, and the limitations of visual features extracted by the model, these methods often struggle to accurately distinguish between them, resulting in classification errors.

Figure 12. Albatross and white-necked duck misclassified images.

The experimental results indicate that on the CUB-200-2011 dataset, our proposed algorithm outperforms ProtoNet, Relation, PARN, SAML, DeepEMD, LRPABN, BSNet(D&C), CTX, FRN, FRN+TDM, BiFRN, and FS-FGIC by 1.94 to 11.79 percentage points in 5-way 1-shot accuracy, and by 2.26 to 8.76 percentage points in 5-way 5-shot accuracy. Compared with other fine-grained image classification models, our system achieves an accuracy improvement of 0.93 to 10.75 percentage points in the 5-way 1-shot job and 1.04 to 5.9 percentage points in the 5-way 5-shot task on the Stanford Cars dataset. These outcomes provide a thorough demonstration of our suggested algorithm's performance in fine-grained few-shot picture categorization tasks.

3.4.2. Analysis of Ablation Study

To further investigate the actual impact of the DRFF Module and MRA Block on classification accuracy, detailed ablation experiments were conducted on the three benchmark datasets using Conv4 and ResNet-12 as feature extractors, respectively. The results of the ablation study are presented in Tables 5–7.

Table 5. Quantitative comparison of ablation experiments on CUB-200-2011 dataset.

Backbone	Method	CUB	
		1-Shot	5-Shot
Conv4	Baseline(ProtoNet)	64.82 ± 0.23	85.74 ± 0.14
	A	80.11 ± 0.20	93.20 ± 0.10
	B	80.39 ± 0.19	93.42 ± 0.11
	A + B	81.83 ± 0.19	94.65 ± 0.10
ResNet-12	Baseline(ProtoNet)	81.02 ± 0.20	91.93 ± 0.11
	A	85.94 ± 0.18	95.53 ± 0.09
	B	85.72 ± 0.18	94.96 ± 0.08
	A + B	87.38 ± 0.18	96.99 ± 0.09

Table 6. Quantitative comparison of ablation experiments on Stanford Cars dataset.

Backbone	Method	Cars	
		1-Shot	5-Shot
Conv4	Baseline(ProtoNet)	50.88 ± 0.23	74.89 ± 0.18
	A	75.91 ± 0.20	91.66 ± 0.09
	B	75.99 ± 0.20	90.61 ± 0.10
	A + B	76.32 ± 0.20	91.84 ± 0.09
ResNet-12	Baseline(ProtoNet)	85.46 ± 0.19	95.08 ± 0.08
	A	91.03 ± 0.18	98.06 ± 0.08
	B	90.97 ± 0.16	97.99 ± 0.09
	A + B	91.37 ± 0.18	98.53 ± 0.08

Table 7. Quantitative comparison of ablation experiments on Stanford Dogs dataset.

Backbone	Method	Dogs	
		1-Shot	5-Shot
Conv4	Baseline(ProtoNet)	46.66 ± 0.21	70.77 ± 0.16
	A	62.91 ± 0.20	82.66 ± 0.09
	B	63.99 ± 0.20	81.61 ± 0.10
	A + B	64.69 ± 0.20	83.14 ± 0.09
ResNet-12	Baseline(ProtoNet)	73.81 ± 0.21	87.39 ± 0.12
	A	76.03 ± 0.18	89.06 ± 0.08
	B	76.97 ± 0.16	88.29 ± 0.09
	A + B	77.98 ± 0.18	89.78 ± 0.08

To validate the synergistic effect of the MRA Block and the DRFF Module in fine-grained image classification tasks, four sets of experiments were conducted. In Table 1, 'A' refers to using only the MRA Block without the DRFF Module, aiming to enhance feature representation by capturing crucial information in the images. 'B' represents using only the DRFF Module, enhancing the model's discriminative ability by fusing features from different levels. 'A + B' denotes the complete model proposed in this paper. When both modules A and B are used simultaneously, the model achieves accuracies of 87.38% and 96.99% for the 5-way 1-shot and 5-way 5-shot tasks, respectively, on the CUB-200 dataset. Therefore, the experimental results demonstrate that there is no mutual exclusion between the two modules, indicating strong compatibility, and they can be used together to improve the classification accuracy of fine-grained images.

3.5. Visualized Analysis

In this study, the Grad-CAM [39] class activation visualization method was employed to experimentally test five randomly selected images from the CUB-200-2011 dataset, aiming to validate the learned feature distribution of the model. Figure 13 visually presents the comparison of the visualization results between our method and ResNet-12 in the small-sample image classification task. In the figure, the first row shows the original images, the second row displays the heatmaps generated by the ResNet-12 model, and the third row illustrates the heatmaps generated by our model. Through comparison, it is evident that when using the ResNet-12 network to extract image features, the generated heatmaps exhibit relatively scattered focus on discriminative regions, lacking concentration. In contrast, the heatmaps generated by our model can more accurately localize discriminative regions, such as the wings of birds and other critical parts. By applying the class activation visualization method, the effectiveness of our model in learning feature distributions and discriminative regions in fine-grained small-sample image classification tasks is further validated.

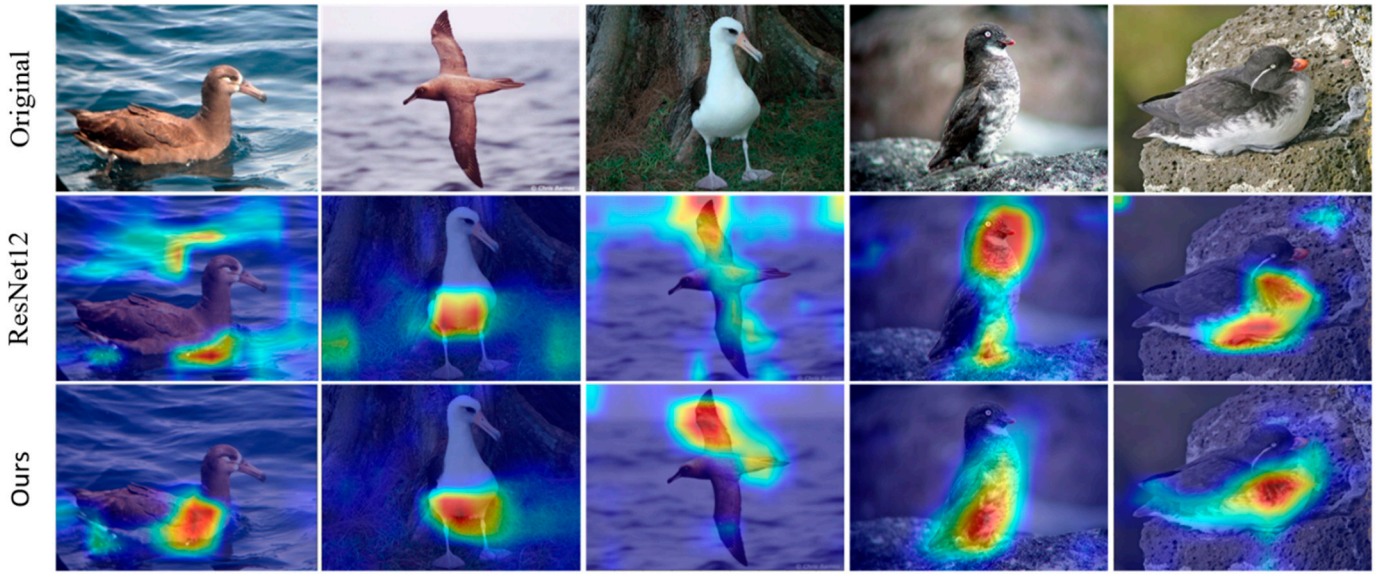


Figure 13. Visual analysis renderings.

4. Discussion

We propose a bidirectional feature reconstruction method for fine-grained few-shot image classification tasks. Experiments and ablation studies show that the MRA Block and DRFF modules improve the model's sensitivity to subtle inter-category differences by combining channel attention and window self-attention. They also enhance the model's adaptability to inter-class and intra-class variations through bidirectional feature reconstruction at different scales. This is crucial for addressing the poor robustness of feature vectors extracted by the network, caused by significant intra-class differences and minimal inter-class differences in fine-grained few-shot images.

Despite the promising results of our algorithm on various few-shot datasets, it has several limitations that need to be addressed in future work. Firstly, the algorithm's performance heavily relies on the choice of the embedding backbone, especially pre-trained models. While these pre-trained models provide strong support for few-shot learning algorithms due to their excellent feature representations, their heavy reliance in the algorithm is a matter that warrants further investigation in future studies. Secondly, although we have introduced skip connections in the modules to enhance the information retention capability of the model, the complexity of the model may still limit its potential performance improvements. Many few-shot methods adopt simple pipelines but perform well on few-shot datasets. We plan to continue in this research direction.

Furthermore, we need to further explore our improvements in future research, particularly when the training and testing sets come from different domains. We believe that more efficient designs for the model have yet to be discovered and that the balance between hard and easy samples has not been fully studied. We are committed to continuing this line of research and will share our findings in a timely manner. Our work is closely related to other studies in the field, such as those by Zheng et al. [40], Yang et al. [41], and Wang et al. [42]. These studies not only demonstrate the potential of machine learning in various fields but also provide insights into multi-scale feature representation and fusion, which align with the strategies adopted in our method.

We suggest that future research based on our work should consider more comprehensive and broader comparisons. Additionally, the development of fine-grained few-shot image classification technology can advance fields such as biodiversity monitoring and medical image analysis. In biodiversity monitoring, our model could assist in identifying and classifying different species, thereby providing technical support for ecological conservation and biodiversity research. In the field of medical image analysis, fine-grained image classification capabilities can be used to improve the accuracy and efficiency of disease

diagnosis, especially in applications requiring the identification of subtle pathological changes. We leave these limitations and discussions for future studies.

5. Conclusions

This study introduces a novel fine-grained image classification algorithm designed to tackle the issues of large intra-class variation and small inter-class variation. The approach uses a bidirectional feature reconstruction mechanism, combining a Mixed Residual Attention (MRA) Block and a Dual-Reconstruction Feature Fusion (DRFF) Module. The MRA Block merges channel attention with window-based self-attention to capture local details and improve feature representation. Meanwhile, the DRFF module enhances cross-layer feature fusion and adjusts to both inter-class and intra-class variations, boosting the model's sensitivity to subtle category differences.

We conducted experiments on three benchmark datasets. Specifically, in the 5-way 5-shot task, our method achieved classification accuracies of 96.99%, 98.53%, and 89.78%. These results demonstrate the effectiveness of our approach in handling fine-grained classification tasks with limited samples. This high accuracy not only validates the robustness of our algorithm but also highlights its potential in real-world applications.

Additionally, ablation studies confirmed the utility of each component of our method. The combination of the MRA Block and the DRFF Module showed a synergistic effect, leading to significantly improved classification accuracy compared to baseline models. Grad-CAM visualizations further demonstrated our model's ability to focus on discriminative regions, such as the wings of birds, which are crucial for fine-grained classification. Overall, our work makes significant progress in the field of few-shot learning for fine-grained image classification. The proposed method can learn from very few examples and accurately classify fine-grained categories, which is important for applications in fields such as biodiversity monitoring. Future work will explore the scalability of our method to a broader range of categories and investigate its applicability in other areas of computer vision.

Author Contributions: Conceptualization, W.Z.; methodology, W.Z.; software, W.Z.; validation, W.Z.; formal analysis, W.Z.; investigation, W.Z.; resources, W.Z.; data curation, W.Z.; writing—original draft preparation, W.Z.; writing—review and editing, S.L., W.Z., and B.G.; visualization, F.G.; supervision, S.L., W.Z., F.G., J.C., and B.G.; project administration, S.L., W.Z., and B.G.; funding acquisition, B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Philosophy and Social Sciences Planning Project of Tianjin (TJGL19XSX-045).

Data Availability Statement: The data presented in this study can be requested from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zheng, P.; Chen, H.; Hu, S.; Zhu, B.; Hu, J.; Lin, C.S.; Wu, X.; Lyu, S.; Huang, G.; Wang, X. Few-shot learning for misinformation detection based on contrastive models. *Electronics* **2024**, *13*, 799. [CrossRef]
2. Valero-Mas, J.J.; Gallego, A.J.; Rico-Juan, J.R. An overview of ensemble and feature learning in few-shot image classification using siamese networks. *Multimed. Tools Appl.* **2024**, *83*, 19929–19952. [CrossRef]
3. Yaqing, W.; Quanming, Y.; Kwok James, T.; Ni Lionel, M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **2020**, *53*, 1–34.
4. Zhu, Y.; Liu, C.; Jiang, S. Multi-attention Meta Learning for Few-shot Fine-grained Image Recognition. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-2020), Yokohama, Japan, 7–15 January 2021; pp. 1090–1096.
5. Wertheimer, D.; Tang, L.; Hariharan, B. Few-shot classification with feature map reconstruction networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, Nashville, TN, USA, 20–25 June 2021; pp. 8012–8021.
6. Doersch, C.; Gupta, A.; Zisserman, A. Crosstransformers: Spatially-aware few-shot transfer. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21981–21993.
7. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4077–4087.

8. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-UCSD Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
9. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–20 June 2020; pp. 12203–12213.
10. Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; Xue, J.H. BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Trans. Image Process.* **2020**, *30*, 1318–1331. [CrossRef]
11. Kim, J.; Kim, T.; Kim, S.; Yoo, C.D. Edge-labeling graph neural network for few-shot learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11–20.
12. Tang, Z.; Yang, H.; Chen, C.Y.C. Weakly supervised posture mining for fine-grained classification. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23735–23744.
13. Chang, D.; Tong, Y.; Du, R.; Hospedales, T.; Song, Y.Z.; Ma, Z. An erudite fine-grained visual classification model. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7268–7277.
14. Lyu, Y.; Jing, L.; Wang, J.; Guo, M.; Wang, X.; Yu, J. Siamese transformer with hierarchical concept embedding for fine-grained image recognition. *Sci. China Inf. Sci.* **2023**, *66*, 132107. [CrossRef]
15. Wu, J.; Chang, D.; Sain, A.; Li, X.; Ma, Z.; Cao, J.; Guo, J.; Song, Y.Z. Bi-directional feature reconstruction network for fine-grained few-shot image classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 2821–2829.
16. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22367–22377.
17. Xing, E.; Jordan, M.; Russell, S.J.; Ng, A. Distance metric learning with application to clustering with side-information. *Adv. Neural Inf. Process. Syst.* **2002**, *15*, 505–512.
18. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, 2017, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
19. Rusu, A.A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; Hadsell, R. Meta-learning with latent embedding optimization. *arXiv* **2018**, arXiv:1807.05960.
20. 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 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
21. Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; Luo, J. Revisiting local descriptor based image-to-class measure for few-shot learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7260–7268.
22. Min, W.; Wang, Z.; Liu, Y.; Luo, M.; Kang, L.; Wei, X.; Wei, X.; Jiang, S. Large scale visual food recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9932–9949. [CrossRef] [PubMed]
23. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12581–12600. [CrossRef]
24. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Virtual Conference, 11–17 October 2021; pp. 22–31.
25. Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; Girshick, R. Early convolutions help transformers see better. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 30392–30400.
26. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating convolution designs into visual transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Virtual Conference, 11–17 October 2021; pp. 579–588.
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–20 June 2020; pp. 11534–11542.
28. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Virtual Conference, 11–17 October 2021; pp. 1833–1844.
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Virtual Conference, 11–17 October 2021; pp. 10012–10022.
30. Patel, K.; Bur, A.M.; Li, F.; Wang, G. Aggregating global features into local vision transformer. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 1141–1147.
31. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.

32. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. Resnest: Split-attention networks. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2736–2746.
33. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 3560–3569.
34. Wu, Z.; Li, Y.; Guo, L.; Jia, K. Parn: Position-aware relation networks for few-shot learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6659–6667.
35. Hao, F.; He, F.; Cheng, J.; Wang, L.; Cao, J.; Tao, D. Collect and select: Semantic alignment metric learning for few-shot learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8460–8469.
36. Huang, H.; Zhang, J.; Zhang, J.; Xu, J.; Wu, Q. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *IEEE Trans. Multimed.* **2020**, *23*, 1666–1680. [CrossRef]
37. Lee, S.; Moon, W.; Heo, J.P. Task discrepancy maximization for fine-grained few-shot classification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5331–5340.
38. Ma, Z.X.; Chen, Z.D.; Zhao, L.J.; Zhang, Z.C.; Luo, X.; Xu, X.S. Cross-Layer and Cross-Sample Feature Optimization Network for Few-Shot Fine-Grained Image Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4136–4144.
39. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
40. Zheng, Q.; Zhao, P.; Wang, H.; Elhanashi, A.; Saponara, S. Fine-grained modulation classification using multi-scale radio transformer with dual-channel representation. *IEEE Commun. Lett.* **2022**, *26*, 1298–1302. [CrossRef]
41. Yang, M.; Bai, X.; Wang, L.; Zhou, F. HENC: Hierarchical embedding network with center calibration for few-shot fine-grained SAR target classification. *IEEE Trans. Image Process.* **2023**, *32*, 3324–3337. [CrossRef]
42. Wang, Y.; Ji, Y.; Wang, W.; Wang, B. Bi-channel attention meta learning for few-shot fine-grained image recognition. *Expert Syst. Appl.* **2024**, *242*, 122741. [CrossRef]