International Journal of
Network Dynamics and Intelligence (IJNDI)

Scilight

Article

Few-Shot Classification Using Ensemble of Multi-Scale Median-Enhanced Features

Chao Yang, Sunjie Zhang *, and Zhanqiang Liu

School of Control Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
* Correspondence: zhang_sunjie@126.com

Received: 3 July 2025; Revised: 25 November 2025; Accepted: 28 December 2025; Published: 9 March 2026

Abstract: Few-shot learning aims to train classifiers with limited samples for novel object recognition, facing key challenges in feature extraction robustness and discriminative representation. To address these issues, we propose a Median-Enhanced Multi-Scale Adaptive Network. Firstly, an adaptive fusion convolution module with deformable kernels is designed to capture spatially transformed features, improving cross-domain adaptability. Next, a median-enhanced attention mechanism integrates median filtering with channel attention, effectively suppressing feature noise and outliers while highlighting discriminative patterns. Finally, we develop a hierarchical metric learning framework that combines multi-scale feature representations with learnable similarity metrics. Experimental results demonstrate that the proposed method outperforms state-of-the-art approaches, achieving accuracy gains of 1.27% (1-shot) and 1.12% (5-shot) on Mini-ImageNet, 1.76%/1.52% on Tiered-ImageNet, and 2.28%/2.21% on CUB, compared to the SetFeat model.

Keywords: few-shot learning; dynamic convolution; deformable convolution; fusion attention mechanism; median enhancement

1. Introduction

Few-shot learning (FSL) enables models to recognize new categories with very few training examples, addressing the challenge of transferring knowledge from seen to unseen classes, especially in scenarios where acquiring labeled data is expensive or impractical. Early FSL approaches, such as those relying on generative models (e.g., variational autoencoders) and inference models [1], faced challenges in terms of scalability and generalization to complex data distributions. With the rise of deep learning, the focus has shifted to three key strategies: meta-learning, data augmentation, and metric learning.

Meta-learning methods, such as Model-Agnostic Meta-Learning [2] (MAML), emphasize the ability to adapt quickly to new tasks by learning a meta-learner that can optimize models efficiently with limited data. More recent meta-learning approaches, such as MetaOptNet [3], leverage differentiable convex optimization to adapt to novel tasks, but struggle with complex feature distributions—particularly when input data contains high variability or noise, as the convex formulation constrains flexibility in capturing non-linear patterns. Data augmentation alleviates data scarcity through various techniques, such as clustering descriptors and background mixing for visual data and cross-modal embeddings for semantic data. Notable among these is MixtFSL [4], which models feature spaces as mixtures of components to capture category diversity, but its scalability is limited in high-noise scenarios where data distributions are skewed or ambiguous. Despite their potential, both approaches struggle with feature representation robustness and generalization, especially in complex real-world scenarios, which has led to the dominance of metric learning.

Metric learning [5], now the dominant approach in FSL, classifies samples based on similarity measurements. Early methods, such as Siamese Networks and Prototypical Networks [6], relied on distance-based comparisons, while Matching Networks introduced attention-based memory and Relation Networks used nonlinear classifiers to enhance query-class relationships. More recent advancements, including Deep KNN and attention-enhanced techniques, improve spatial relationships through mechanisms such as positional encoding. Contemporary metric learning methods have made significant strides: DeepEMD [7] employs the Earth Mover’s Distance to measure fine-grained feature similarities, capturing subtle differences between samples, but its computational complexity increases sharply with feature dimensionality, limiting efficiency in multi-scale scenarios. DMF [8] introduces dynamic alignment via meta-filters to adapt to task-specific features, yet it overlooks cross-scale feature interactions, leading to suboptimal performance when objects exhibit varying sizes or resolutions. SetFeat [9] has further refined metric learning by representing images as sets of feature vectors rather than single embeddings, allowing for richer, more flexible representations. However, SetFeat still struggles with information loss during feature aggregation and limited filter diversity, which reduces its discriminative power.

Among existing metric learning methods, several representative approaches have inspired our work. SetFeat [9] laid a critical foundation for set-based feature representation but suffered from two key limitations: (1) information loss during feature aggregation, where valuable local details are smoothed out in global pooling; (2) limited diversity in convolutional filters, restricting its ability to capture complex spatial transformations. These weaknesses motivated us to design two core modules: the Adaptive Fusion Convolution Module (to enhance filter diversity) and the Median-Enhanced Attention Module (to preserve local details while suppressing noise). Additionally, MELR [10] demonstrated the effectiveness of graph neural networks in improving generalization, while DeepEMD [7] highlighted the importance of fine-grained similarity metrics—insights we integrated into our hierarchical metric learning framework.

Our Adaptive Fusion Convolution Module also draws inspiration from two advanced convolution techniques: dynamic convolution [8] and deformable convolution. Dynamic convolution's ability to generate task-specific kernels based on input features inspired our design of adaptive kernel fusion, while deformable convolution's flexible spatial sampling (via learned offsets) informed our approach to capturing non-rigid object transformations. By combining these two mechanisms in parallel, we address a critical limitation of prior single-convolution methods: poor adaptability to cross-domain variations (e.g., changes in object pose or background clutter).

Attention mechanisms [11–13] have significantly enhanced metric learning by dynamically weighting key features and suppressing irrelevant or noisy information, improving the discriminative power of the learned representations. Cross-attention frameworks improve multimodal alignment, while task-specific models like TDM improve fine-grained classification. In medical image analysis, attention mechanisms have been particularly valuable in refining feature selection and focusing on critical image regions [14]. Despite these advancements, attention-based methods face challenges such as sensitivity to noisy or outlier features, high computational complexity, and risks of overfitting due to excessive reliance on model parameters [15]. Recent studies have shown that combining attention with graph neural networks (GNNs) [10] can improve generalization by leveraging graph structures to compensate for limited data.

Prototype-based methods have also made significant strides in FSL. For instance, prototype-assisted contrastive adversarial networks have addressed weak-shot learning challenges [16], especially in scenarios with weakly labeled data. Similarly, adversarial self-attention networks have been applied to cross-domain FSL tasks, such as pipeline fault diagnosis, by aligning subdomains for improved data augmentation and model robustness [17].

Despite these advancements, challenges remain in improving the robustness of attention mechanisms and their ability to handle diverse, noisy data. Future research focused on refining attention techniques, especially through multi-task learning and domain adaptation, will be crucial to advancing FSL across various domains. To overcome these challenges, this paper introduces a novel approach that integrates adaptive fusion convolution with a median-enhanced attention module. The main contributions include:

(1) Adaptive Fusion Convolution Module: Combines parallel dynamic and deformable convolutions to effectively capture spatially transformed features and enhance adaptability to diverse data.

(2) Median-Enhanced Attention Module: Integrates median pooling with average and max pooling to suppress noise and outliers, improving robustness and feature extraction.

(3) Enhanced SetFeat Framework: Focuses on improving feature maps and strengthens the set-based metric's performance. Experimental results on TieredImageNet and CUB datasets demonstrate the effectiveness of the proposed approach.

2. Adaptive Fusion Median Attention Algorithm

2.1. Problem Definition

In few-shot learning, the goal is to accurately classify new, unlabeled samples from novel categories using only a limited number of training samples. The data is typically divided into three sets: the support set S, the query set Q, and the auxiliary set A. The auxiliary set consists of base classes, while the support and query sets correspond to the new classes. Importantly, the label space of the auxiliary set does not overlap with that of the new classes, ensuring that the model learns to transfer knowledge rather than simply memorize labels. Given a support set S containing n categories, each with k samples, the support set can be represented as $S = \{s_1, s_2, \ldots, s_n\}$, where $n = K \times N$, meaning the support set consists of Q categories, each with K images. The query set Q contains m unlabeled samples, represented as $Q = \{q_1, q_2, \ldots, q_m\}$, which need to be classified into one of the N categories in the support set. This setup is known as the N-way K-shot classification task.

During training, the model learns through multiple episodes, each involving a support set S and a query set Q. These episodes simulate the N-way K-shot classification scenario, with the aim of enabling the model to effectively learn from only a few samples. In each episode, the support and query sets are randomly sampled from the auxiliary set A. Through this episodic training approach, the model learns to extract category feature representations and develop discriminative abilities, which better prepares it for few-shot classification tasks during the testing phase.

2.2. Algorithm Overview

The small-sample algorithm proposed in this paper is illustrated in Figure 1. The model framework consists of three key components: the Adaptive Fusion Convolutional Mod ule, the Median Enhancement Attention Module, and the Ensemble Feature Extraction Module.

The Adaptive Fusion Convolutional Module improves the model's adaptability to di verse data by capturing features with spatial transformations across different receptive fields, thus enhancing its ability to understand the overall layout of the input data.

The Median Enhancement Attention Module introduces median pooling to increase robustness. This module captures spatial features at various scales and generates a spatial attention map, helping the model focus on relevant areas.

Finally, the Ensemble Feature Extraction Module processes the features to produce an ensemble feature map. This map is then compared using a minimum sum metric, which measures the similarity between the extracted features and generates the final output.

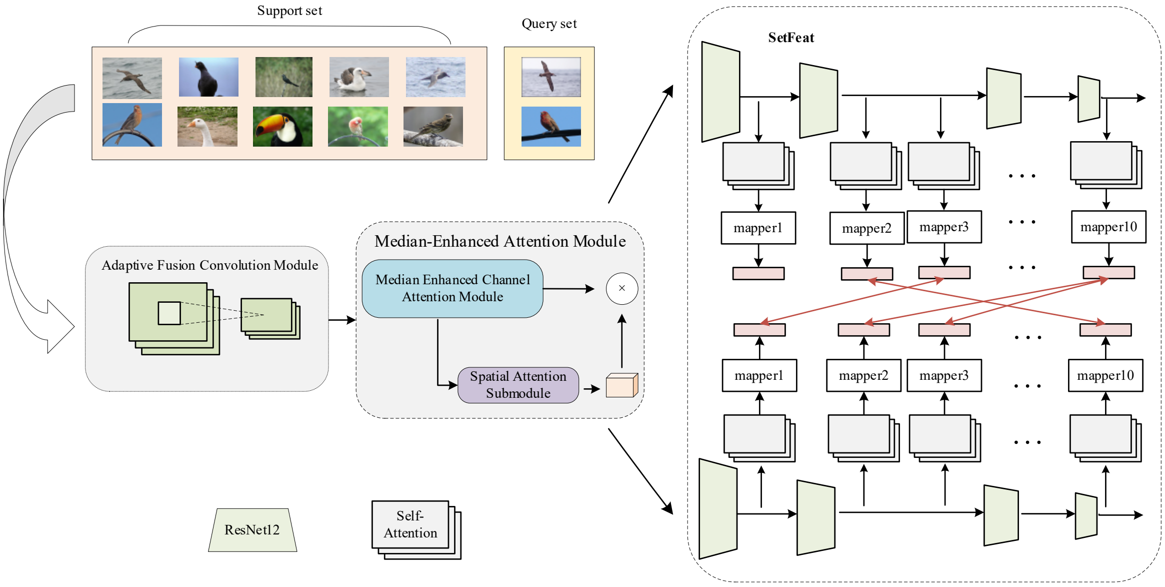
Support set
Query set
SetFeat
mapper1
mapper2
mapper3
. . .
mapper10
Adaptive Fusion Convolution Module
Median-Enhanced Attention Module
Median Enhanced Channel Attention Module
×
Spatial Attention Submodule
ResNet12
Self-Attention

Figure 1. Overall framework diagram.

2.3. Adaptive Fusion Convolutional Module

To address challenges in small-sample image classification, such as poor generalization and insufficient feature extraction due to limited training data, this paper proposes a new module that combines dynamic convolution with deformable convolution in parallel. The structure of the adaptive fusion convolution module is shown in Figure 2. Dynamic convolution adaptively generates convolutional kernels based on input features, enhancing the model's ability to capture relevant features across different samples [8]. Deformable convolution, on the other hand, improves the model's ability to handle deformations, complex backgrounds, and irregularly structured objects by dynamically adjusting the sampling positions of the kernels [7]. This parallel module enhances the model's feature extraction capacity and adaptability, significantly improving classification accuracy and reducing overfitting in small-sample conditions.

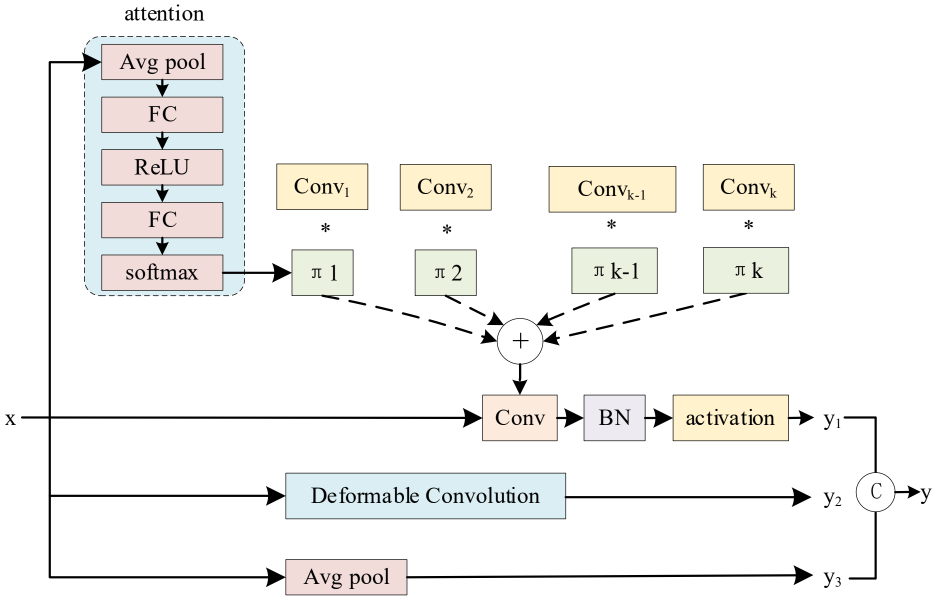

Figure 2. Adaptive fusion convolution module.

Unlike traditional convolution, dynamic convolution generates *k* different convolutional kernels for each input channel, maintaining consistency in scale and channel count. These kernels are followed by batch normalization and ReLU activation. The kernels are then fused using attention weights, which produce the final kernel parameters for the layer. In the attention layer, global average pooling extracts global spatial features, which are mapped to *k* dimensions through two fully connected layers, followed by Softmax normalization. The resulting attention weights dynamically allocate the convolutional kernels, allowing the model to adapt the fixed convolutional kernel to the weights {$\pi_1, \ldots, \pi_k$}, improving the model's feature extraction and expression capabilities.

Deformable convolution is an extension of regular convolution, as shown in Figure 3 It improves on traditional convolution by adjusting the grid sampling positions through learned offsets, enabling the convolutional region to focus on areas of interest as needed. In traditional convolution, the network operates with a fixed-size square receptive field, regardless of the input image's size or shape. This lack of flexibility limits the ability to capture global information during feature extraction. The output y of traditional 2D convolution is calculated as shown in Equation (1).

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{1}$$

In traditional convolution, *w* represents the weights of the convolutional kernel, and *R* denotes the set of positions, including p_0 and other positions p_n within the neighborhood.

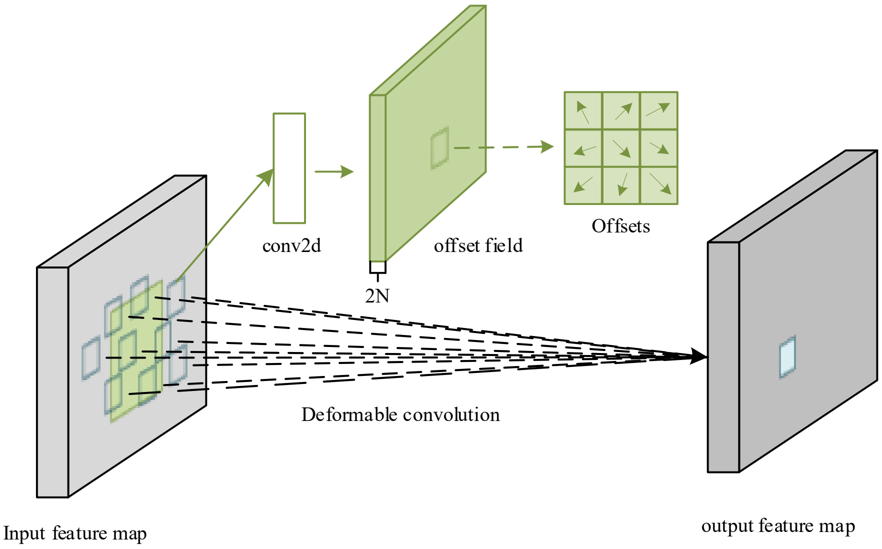

Figure 3. Deformable convolution.

In deformable convolution, the regular grid R is modified by learned offsets, as shown in Equation (2).

$$y(p)=\sum_{k=1}^{K} w_k \cdot x(p+p_k+\Delta p_k)\cdot \Delta m_k \tag{2}$$

In this context, p represents the original position of the convolutional kernel, p_k denotes other positions within the neighborhood of each standard convolutional kernel, and Δp_k is the offset applied to displace the kernel. The modulation factor Δm_k ranges between [0,1]. During feature extraction, if the offset sampling points stray from the ideal area, they may fall into regions unrelated to the object of interest. To address this, the modulation factor Δm_k is introduced as a control mechanism, ensuring that sampling points remain focused on relevant features, thereby improving the effectiveness of feature extraction.

2.4. Median-Enhanced Attention Module

Median enhancement is typically achieved through a median filter, which replaces each pixel with the median value of its local neighborhood. This operation reduces noise while preserving important details [18]. As a classic nonlinear filtering technique, median pooling is primarily used to eliminate noise. In deep learning, spatial channel attention mechanisms help the model focus on important spatial and channel features. By combining median pooling with attention mechanisms, we can leverage the noise reduction capabilities of median pooling along with the dynamic feature selection provided by attention, enhancing the model's robustness to noise and its ability to focus on key features.

Existing hybrid pooling attention mechanisms (e.g., CBAM [11]) typically combine average and max pooling to capture global intensity trends and local peak responses, respectively. However, these methods exhibit limitations in handling noisy or outlier-rich data: average pooling is easily skewed by extreme values (e.g., bright spots in images), while max pooling overemphasizes isolated strong responses, often overlooking broader contextual patterns. In contrast, our Median-Enhanced Attention introduces median pooling as a third component, which inherently resists outliers by selecting the middle value in a local neighborhood—effectively suppressing salt-and-pepper noise or anomalous features while preserving structural details. The structure of the median-enhanced attention module is shown in Figure 4.

This triple-pooling strategy (average + max + median) enables the model to:

(1) Suppress salt-and-pepper noise and outliers (via median pooling), which is critical for small-sample scenarios where data quality is often compromised;
(2) Capture global intensity trends (via average pooling) to retain overall feature distributions;
(3) Highlight discriminative local regions (via max pooling) such as object edges or textures.

The median, as a robust statistical measure, effectively resists the influence of outliers, leading to more stable and reliable feature representations. Given a set of elements $\{x_1, x_2, \ldots, x_{k^2}\}$ in a pooling window R, the output M of median pooling can be defined as:

$$M=\begin{cases} x_{\left(\frac{k^2+1}{2}\right)} & \text{if } k^2 \text{ is odd} \\ \dfrac{x_{\left(\frac{k^2}{2}\right)}+x_{\left(\frac{k^2}{2}+1\right)}}{2} & \text{if } k^2 \text{ is even} \end{cases} \tag{3}$$

Unlike the mean, the median is less sensitive to noise and extreme values. Traditional channel attention mechanisms often use global average pooling and max pooling, which are effective but may lack robustness when handling noisy data or outliers. Average pooling can be overly influenced by extreme values, while max pooling tends to focus on the strongest local response, potentially overlooking global context. Introducing median pooling addresses these issues by providing effective noise suppression. Attention mechanisms can then adaptively adjust their focus based on the context within the image or feature map. By integrating median pooling into the spatial channel attention module, the network gains enhanced denoising capabilities in noisy regions and better detail preservation in areas with rich features. This enables adaptive noise handling and improved feature enhancement. The three pooling formulas are as follows:

$$F_{\text{avg}}(c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F(i,j,c) \tag{4}$$

$$F_{\max}(c)=\max_{1\le i\le H,1\le j\le W}F(i,j,c) \tag{5}$$

$$F_{\text{med}}(c) = \underset{1\le i\le H, 1\le j\le W}{\text{median}} F(i,j,c) \tag{6}$$

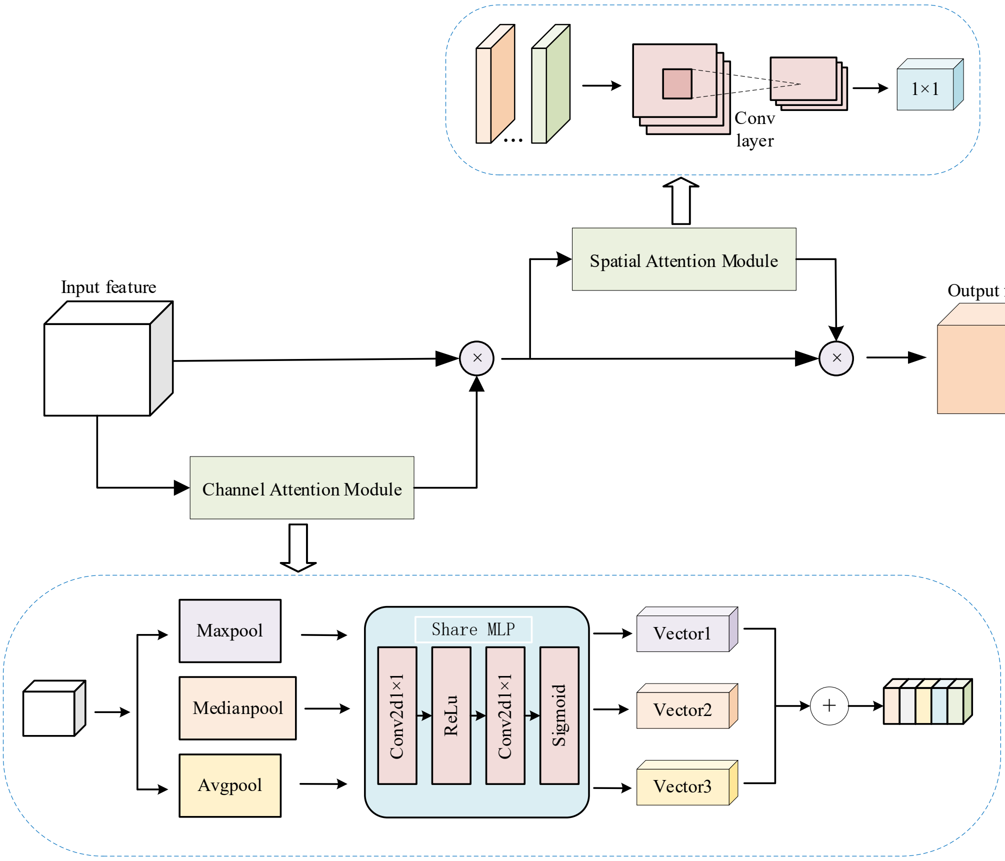

Figure 4. Median-enhanced attention module.

The input feature map undergoes three types of global pooling: average pooling (Avg Pool), max pooling (Max Pool), and median pooling (Median Pool). Each pooling operation results in a feature map of size 1 × 1 × *C*, where *C* is the number of channels. These pooling results are then passed through a shared multi-layer perceptron (MLP) consisting of two 1 × 1 convolutional layers and a ReLU activation function. The first convolutional layer reduces the feature dimension from *C* to *C*/*r*, where *r* is the reduction ratio, and the second layer restores the dimension to *C*. A Sigmoid activation function is then applied to scale the output values to the range [0,1], producing three attention maps. These attention maps are combined element-wise to form the final channel attention map. Equations (7) and (8) are as follows:

$$F_c = \sigma(\text{MLP}(\mathbf{F}_{\text{avg}})) + \sigma(\text{MLP}(\mathbf{F}_{\text{max}})) + \sigma(\text{MLP}(\mathbf{F}_{\text{med}})) \tag{7}$$

$$F' = F_c \odot F \tag{8}$$

where σ denotes the Sigmoid function and ⊙ represents element-wise multiplication.

To improve the model's performance in complex scenes, this paper introduces multi-scale depthwise convolutions in the spatial attention submodule. Initially, a 5 × 5 convolutional layer extracts basic features, which are then passed through several depthwise convolutional layers of varying sizes to capture multi-scale features. These features are combined element-wise to form a fused feature map, which is subsequently processed by a 1 × 1 convolutional layer to produce the final spatial attention map. Finally, this attention map is multiplied element-wise with the channel-weighted feature map to generate the final output. Equations (9) and (10) are as follows:

$$F_s = \sum_{i=1}^{n} D_i(F') \tag{9}$$

$$F'' = \text{Conv}_{1\times1}(F_s) \odot F' \tag{10}$$

where D_i denotes the depthwise convolution operations of varying sizes, and *n* represents the number of such operations.

2.5. Set-feature Extraction Module

The Set-feature extraction module aims to map the input features x to an ensemble feature *H*. This paper integrates separate self-attention mappers within each ResNet layer, as shown in Figure 5a. Each mapper consists of a single attention head, avoiding the need for fully connected layers to concatenate multi-head outputs. The feature map pers operate independently, with each one extracting its own distinct set of features. The details of the *m*-th mapper are shown in Figure 5b. The resulting feature vector $\mathbf{z}_b$ is linearly transformed to obtain the query θ_m^q, key θ_m^k, and value θ_m^v representations. The attention scores are then computed by taking the dot product between the query and key, followed by scaling and normalization using the Softmax function:

$$\beta_m = \text{Softmax}\left(q(\mathbf{z}_{b_m} \mid \theta_m^q)k(\mathbf{z}_{b_m} \mid \theta_m^k)^{\cdot} \ / \sqrt{d_k}\right) \tag{11}$$

Among them, β_m is the attention power coefficient on block $\mathbf{z}_m$, and $\sqrt{d_k}$ is a scaling factor. Then, each value is multiplied by the corresponding attention power and summed, yielding the output, which has the following form:

$$\mathbf{a}_m = \beta_m v(\mathbf{z}_{b_m} \mid \theta_m^v) \tag{12}$$

This paper introduces a residual connection following the attention calculation. When the dimensions do not align, a 1 × 1 convolution with a stride and kernel size similar to down sampling is applied. Finally, average pooling is used to generate the feature vector hm.

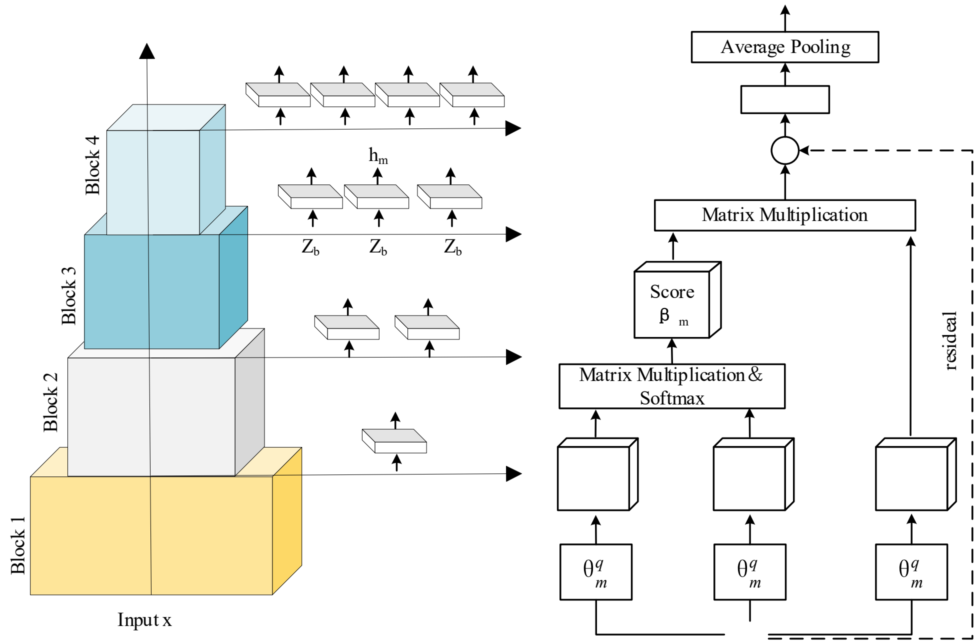

Figure 5. Set feature extraction module: (**a**) Overall structure of the set-feature extraction module; (**b**) Detailed workflow of the attention-based feature mapping and aggregation.

2.6. Ensemble Metric and Training Process

To classify the query, its feature set is compared with the feature sets of each instance in the support set for each class. This enables the inference of the query's category. To facilitate distance-based methods in prototype networks, an inter-set measure is required to compute distances between sets. This paper defines the feature centroid extracted by mapper *m* from the support set *S*, as shown in Equation (13):

$$\bar{\mathbf{h}}_m(\mathsf{S}) \equiv \frac{1}{|\mathsf{S}|}\sum_{\mathbf{x}\in\mathsf{S}} \mathbf{h}_m(\mathbf{x}) \tag{13}$$

Here, $\mathbf{h}_m(x)$ denotes the ensemble feature extracted from the support set. The classification is then determined by aggregating the sum of the minimum distances between the query and the centroids of the support set, as calculated by the mapper:

$$\mathrm{d}_{\mathrm{sm}}(\mathbf{x}_{\mathrm{q}},\mathrm{S}^{\mathrm{n}})=\sum_{\mathrm{i}=1}^{\mathrm{M}}\min_{\mathrm{j}=1}^{\mathrm{M}}\mathrm{d}\left(\mathbf{h}_{\mathrm{i}}(\mathbf{x}_{\mathrm{q}}),\overline{\mathbf{h}}_{\mathrm{j}}(\mathrm{S}^{\mathrm{n}})\right) \tag{14}$$

Let $\mathbf{h}_j(S^n)$ represent the feature centroid extracted by mapper m from the support set S^n. This paper employs a negative cosine similarity function, defined as $d(\ ,\ ) = -\cos(\ ,\ )$.

Using the metric d_{set}, we apply the prototype network method with SetFeat to model the probability that a query sample $\mathbf{x}_q$ belongs to class $y=n$ (for an N-class classification task). The softmax function is then used for normalization. Specifically, the probability is computed based on the support set S as follows:

$$p(y=n\,|\,\mathbf{x}_q,\mathrm{S})=\frac{\exp(-d_{\mathrm{set}}(\mathbf{x}_q,\mathrm{S}^n))}{\sum_{\mathrm{S}^i\in\mathrm{S}}\exp(-d_{\mathrm{set}}(\mathbf{x}_q,\mathrm{S}^i))} \tag{15}$$

The training process follows a two-stage approach. In the first stage, a standard pre-training phase is conducted by randomly selecting a batch of instances, $\mathbf{x}$, from the base classes in the training set. During this phase, the paper introduces a fully connected layer, o_m, which transforms the feature representations, $\mathbf{h}_m$, from each mapper into logits for classification across C classes. Each mapper is trained independently using the cross-entropy loss function, as shown in Equation (16).

$$\ell_{\mathrm{pre}}=-\sum_{\mathbf{x}_i\in\mathrm{X}_{\mathrm{batch}}}\sum_{m=1}^{M}\log\frac{\exp(o_{m,y_i}(\mathbf{h}_{m,i}))}{\sum_{c=1}^{C}\exp(o_{m,c}(\mathbf{h}_{m,i}))} \tag{16}$$

Let $o_{m,c}$ denote the output of the fully connected layer in mapper m for class c, and let $\mathbf{h}_{m,i}$ represent the feature embeddings generated by mapper m for input sample $\mathbf{x}_i$, which align with the ground-truth labels of $\mathbf{x}_i$. In the second phase, we transition to episodic training [19] to emulate real-world few-shot learning conditions. This involves removing the fully connected layer introduced in the first stage and constructing episodic tasks from the base dataset. For each task, we randomly select N classes, sampling K-shot and Q-query per class. The softmax function is applied to compute probability distributions over the query samples, and the model parameters are optimized end-to-end via cross-entropy loss minimization.

3. Experimental Results and Analysis

Dataset and Network Architecture To evaluate the effectiveness and applicability of the proposed model, we conducted validation experiments on three public datasets: miniImageNet, Tiered-ImageNet, and CUB, As shown in Figure 6.

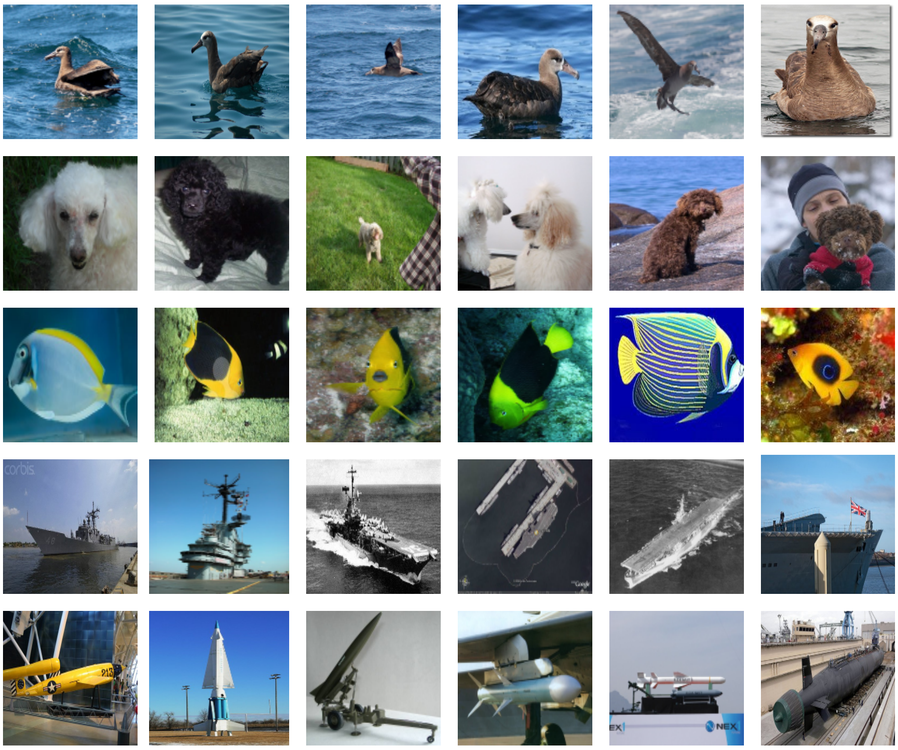

Figure 6. Dataset images.

The mini-ImageNet [20] dataset consists of 100 distinct classes, each containing 600 samples. The dataset is split into training, validation, and test sets in a 64:16:20 ratio, with no overlap between the sets.

Tiered-ImageNet is designed to emphasize the hierarchical relationships among classes. It contains 34 major classes, each with 10 to 30 subclasses, totaling 608 subclasses. The dataset is divided into training, validation, and test sets with a 70:20:10 split. This structure aims to better simulate the cross-class transfer challenge often encountered in few-shot learning.

The CUB200-2011 [21] dataset is a fine-grained dataset introduced in 2010, featuring 11,788 bird images across 200 bird subclasses. The images are labeled with their corresponding class information. Typically, the CUB dataset is divided into training and test sets with a 50:50 split for model evaluation and comparison.

For the experiments, we used an initial learning rate of 0.1, momentum of 0.9, and weight decay of 5×10^{-4} with Nesterov momentum. Normalization and data augmentation techniques were applied throughout the process. During the meta-training phase, Stochastic Gradient Descent (SGD) was employed, and the validation set was used to adjust the optimizer's schedule.

3.1. Comparative Experimental Results

This paper presents comparative experiments on the mini-ImageNet dataset using Conv 4-64 and ResNet-12 as backbone networks. The MELR model enhances the generalization ability of few-shot learning by modeling task relationships through graph neural networks. SetFeat, on the other hand, uses a set-level similarity measurement to improve performance in few-shot image classification. Neg-Margin investigates the impact of negative margins in few-shot classification, proposing a new loss function to enhance model learning effectiveness.

While these models represent some of the most advanced algorithms in the field, the proposed method outperforms them when using Conv-4-64 as the backbone. This demonstrates that the adaptive fusion convolution and median enhancement attention modules are effective in aggregating globally relevant features to local positions, thereby reducing the influence of irrelevant information, as shown in Table 1.

Table 1. Conv-4-64 classification results on the miniImageNet dataset %.

Modle	1-Shot	5-Shot
ProtoNet	49.42 ± 0.78	68.20 ± 0.66
MAML	48.07 ± 1.75	63.15 ± 0.91
RelationNet	50.44 ± 0.82	65.32 ± 0.70
IMP	49.60 ± 0.80	68.10 ± 0.80
MemoryNet	53.37 ± 0.48	66.97 ± 0.35
Neg-Margin	52.84 ± 0.76	70.41 ± 0.66
MixtFSL	52.82 ± 0.63	70.67 ± 0.57
FEAT	55.15 ± 0.20	71.61 ± 0.16
MELR	55.35 ± 0.43	72.27 ± 0.35
BOIL	49.61 ± 0.16	66.45 ± 0.37
SetFeat	57.18 ± 0.89	73.67 ± 0.71
Method (ours)	58.01 ± 0.42	74.92 ± 0.39

The results in Table 2 demonstrate that our model significantly outperforms others in terms of classification accuracy on the mini-ImageNet dataset when using ResNet-12 as the backbone network. Specifically, the classification accuracy for the 5-way 1-shot and 5-way 5-shot tasks improved from 68.32% and 82.71% to 69.59% and 83.83%, respectively, compared to SetFeat. Additionally, our model has surpassed the MELR model, which previously outperformed SetFeat.

Table 2 presents the performance of several leading algorithms on the Tiered-ImageNet dataset using ResNet12 as the backbone network. Distill enhances the model's generalization capability by leveraging both invariant and equivariant representations, achieving a 5-shot accuracy of 87.08%. MixtFSL [13] improves performance by learning multiple mixed components in the feature space to better capture category diversity, reaching an accuracy of 86.16% in the 5-shot setting. The model proposed in this paper outperforms the best-performing SetFeat by 1.76% in the 1-shot task and 1.52% in the 5-shot task. These results further confirm the effectiveness of our approach in few-shot learning tasks.

The evaluation results on the CUB dataset, presented in Table 3, demonstrate a significant performance improvement with the proposed method. The FRN+TDM model enhances discrimination in fine-grained classification tasks by maximizing the differences between tasks, achieving an accuracy of 90.33%. Our approach improves accuracy by approximately 2.07% in the 5-way 1-shot scenario and 2.36% in the 5-way 5-shot scenario

Table 2. Resnet-12 classification results on the miniImageNet dataset %.

Modle	5-Way 1-Shot	5-Way 5-Shot
AdaResNet	56.88 ± 0.62	71.94 ± 0.57
TADAM	58.50 ± 0.30	76.70 ± 0.30
MetaOptNet [16]	62.64 ± 0.61	78.63 ± 0.46
Neg-Margin	63.85 ± 0.76	81.57 ± 0.56
MixtFSL [17]	63.98 ± 0.79	82.04 ± 0.49
Meta-Baseline [18]	63.17 ± 0.23	79.26 ± 0.17
Distill	64.82 ± 0.60	82.14 ± 0.43
DeepEMD [19]	65.91 ± 0.82	82.41 ± 0.56
DMF [20]	67.76 ± 0.46	82.71 ± 0.31
MELR	67.40 ± 0.43	83.40 ± 0.28
SetFeat [5]	68.32 ± 0.62	82.71 ± 0.46
Method (ours)	69.59 ± 0.72	83.83 ± 0.40

Table 3. Few-shot classification results on the Tiered ImageNet dataset %.

Modle	5-Way 1-Shot	5-Way 5-Shot
OptNet [16]	65.99 ± 0.72	81.56 ± 0.53
MTL [21]	65.62 ± 1.80	80.61 ± 0.90
DNS [22]	66.22 ± 0.75	82.79 ± 0.48
Simple	69.74 ± 0.72	84.41 ± 0.55
TapNet	63.08 ± 0.15	80.26 ± 0.12
ProtoNet [23]	68.23 ± 0.23	84.03 ± 0.16
FEAT	70.80 ± 0.23	84.79 ± 0.16
MixtFSL [17]	70.97 ± 1.03	86.16 ± 0.67
DeepEMD [19]	71.16 ± 0.87	86.03 ± 0.58
DMF [20]	71.89 ± 0.52	85.96 ± 0.35
MELR	72.14 ± 0.51	87.01 ± 0.35
Distill	72.21 ± 0.90	87.08 ± 0.58
SetFeat [5]	73.63 ± 0.88	87.59 ± 0.57
Method (ours)	75.39 ± 0.47	89.11 ± 0.58

Compared to FRN+TDM [22]. Additionally, when compared to SetFeat, our model also shows improvement. This enhancement is attributed to several optimizations in our method, including the integration of adaptive fusion convolution and median enhancement attention mechanisms. Specifically, adaptive convolution allows for more detailed and comprehensive feature extraction, capturing positional information from fine-grained feature maps. This refined feature extraction significantly boosts the model's discrimination ability, resulting in higher classification accuracy. Overall, our approach exhibits notable advantages in feature extraction and classification accuracy, underscoring its effectiveness and superiority in few-shot learning tasks.

3.2. Visualization Experiments

To evaluate the effectiveness of the adaptive fusion median attention feature extraction structure, we used the Grad-CAM [18] method for visual analysis of the network's extracted features. The resulting heat maps are shown in Figure 7. Upon comparison, it is evident that the features extracted by ResNet12 and SetFeat are more dispersed and exhibit lower sensitivity to discriminative features. This leads to insufficient feature extraction, which in turn affects classification accuracy, particularly when dealing with tasks involving closely related categories in small-sample datasets, where misclassifications are more common. In contrast, the heatmaps generated by the proposed network structure show more comprehensive and discriminative extraction of image information. The features are more representative of the images, which significantly improves classification performance. This demonstrates the superior capability of the multi-scale variability fusion attention feature extraction structure in enhancing the expressive power of image features.

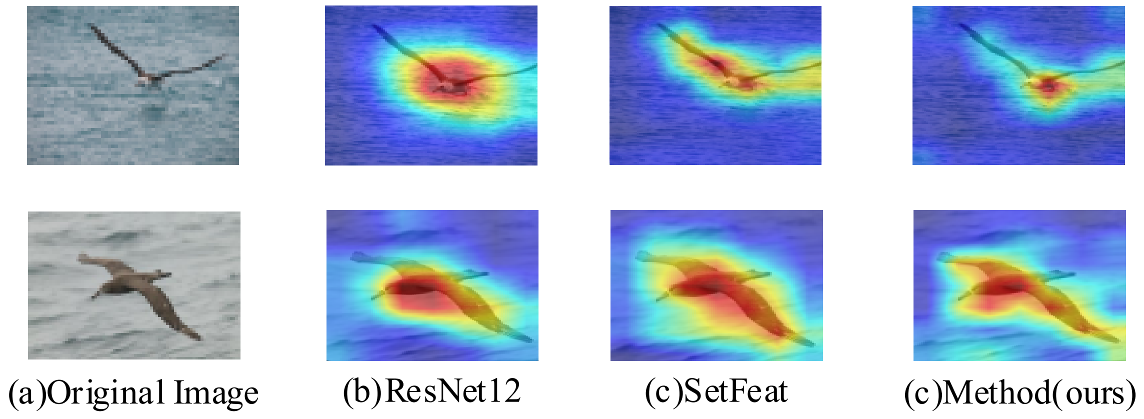

Figure 7. Visual comparison.

3.3. Ablation Experiments

An ablation study was conducted on the CUB dataset using ResNet12 as the backbone to assess the effectiveness of each module in the proposed adaptive fusion median enhancement attention network. The study evaluates how different combinations of modules influence the efficiency of image feature extraction. To identify the optimal configuration for achieving the best few-shot classification performance, we compared the learning outcomes of six distinct structures based on the SetFeat base network, as shown in Table 4. These comparative experiments help clarify the specific contributions of each module to feature extraction and improvements in classification performance. For broader comparison, we also refer to representative few-shot learning methods, including Meta-Baseline [23] and Meta-Transfer Learning [24]. In addition, the comparative results of Adaptive Subspaces [25], Prototypical Networks [26], Relation Networks [27], and LRPABN [28] are reported in Table 4.

Table 4. Few-shot classification results on the CUB dataset %.

Modle	5-Way 1-Shot	5-Way 5-Shot
Matching Nets [15]	45.30 ± 1.03	59.50 ± 1.01
Prototypical Nets [26]	37.36 ± 1.00	45.28 ± 1.03
Relation Nets [27]	53.11 ± 1.01	67.45 ± 0.79
GNN	51.83 ± 0.98	63.69 ± 0.94
DNS [22]	53.15 ± 0.84	81.90 ± 0.60
RCN	66.48 ± 0.88	82.04 ± 0.58
MsKPRN [25]	69.49 ± 0.95	82.94 ± 0.65
Bi-FRN	79.08 ± 0.20	92.22 ± 0.10
LRPABN (TMM2021) [28]	63.63 ± 0.77	76.06 ± 0.58
BSNet(D&C) (TIP2021)	62.84 ± 0.95	85.39 ± 0.56
FRN + TDM (CVPR2022)	76.55 ± 0.21	90.33 ± 0.11
SetFeat [5]	79.60 ± 0.80	90.48 ± 0.40
Method (ours)	81.88 ± 0.78	92.69 ± 0.41

The ablation results in Table 5 reveal the specific contributions of each module to the model's performance:

(1) Adaptive Fusion Convolution Module: When added to the baseline (SetFeat), this module improved 1-shot accuracy by 1.97% (from 79.60% to 81.57%) and 5-shot accuracy by 1.45% (from 90.48% to 91.93%). This outperforms the individual additions of dynamic convolution (1-shot: +1.76%, 5-shot: +1.21%) and deformable convolution (1-shot: +1.82%, 5-shot: +1.26%), indicating that the parallel integration of dynamic and deformable convolutions yields a synergistic effect. This confirms that combining kernel adaptation (dynamic convolution) and flexible spatial sampling (deformable convolution) effectively captures diverse spatial features, especially for objects with varying poses or complex backgrounds.

(2) Median-Enhanced Attention Module: Compared to the baseline with only channel attention (1-shot: 81.07%, 5-shot: 91.36), adding median-enhanced attention achieved higher accuracy (1-shot: 81.28%, 5-shot: 92.03%). The 0.67% improvement in 5-shot accuracy highlights that median pooling—by suppressing noise and outliers—enhances robustness beyond traditional average/max pooling. This is critical for fine-grained datasets like CUB, where subtle differences (e.g., feather patterns) are easily obscured by noise.

(3) Synergistic Effect of Modules: The full model (Method(ours)) outperformed all ablated variants, with 1-shot

accuracy reaching 81.88% (+2.28% compared to baseline) and 5-shot accuracy reaching 92.69% (+2.21% compared to baseline). This indicates that the adaptive fusion convolution module (capturing diverse spatial features) and median-enhanced attention module (suppressing noise) complement each other: improved feature diversity provides richer input for attention, while noise reduction ensures attention focuses on discriminative patterns rather than artifacts.

Table 5. Ablation experiment %.

Modle	5-Way 1-Shot	5-Way 5-Shot
Baseline (SetFeat)	79.60 ± 0.80	90.48 ± 0.40
Baseline + dynamic convolution	81.36 ± 0.76	91.69 ± 0.40
Baseline + deformable convolution	81.42 ± 0.46	91.74 ± 0.53
Baseline + Adaptive Fusion Convolution Module	81.57 ± 0.66	91.93 ± 0.47
Baseline + Channel Attention	81.07 ± 0.79	91.36 ± 0.42
Baseline + Median Enhanced Attention	81.28 ± 0.39	92.03 ± 0.52
Method (ours)	81.88 ± 0.78	92.69 ± 0.41

3.4. Implementation Details

To ensure reproducibility, we provide detailed implementation specifics: Network Architecture: Two backbones were used: Conv4-64 (4 convolutional layers with 64 filters each) and ResNet12 (12 residual blocks with 64/128/256/512 filters). Feature map dimensions at each module: Input images are 3 × 84 × 84 (RGB); after the Adaptive Fusion Convolution Module, output is 64 × 32 × 32; after the Median-Enhanced Attention Module, reduced to 64 × 16 × 16 via stride-2 convolutions; after the Set-feature Extraction Module, the final output is a 512-dimensional ensemble feature vector.

Hyperparameters: Reduction rate $r = 16$ (used in the MLP of channel attention to reduce dimensions from C to C/r before restoring); number of mappers in Set-feature Extraction: $M = 4$; temperature in Softmax for attention scores: $\sqrt{d_\text{k}} = 8$ (where $d_\text{k} = 64$ is the key dimension); dynamic convolution uses $k = 4$ kernels per channel (fused via Softmax); deformable convolution employs 3 × 3 kernels with learned offsets (initialized to 0) and modulation factor $\Delta m_\text{k} \in [0,1]$.

Training Setup: Optimized via SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4} with data augmentation (random cropping 84 × 84→70 × 70, horizontal flipping, color jitter). Episodic training includes 100,000 meta-training episodes (5-way tasks, 1/5 shots) and 600 test episodes (averaged for accuracy).

3.5. Potential Extensions to Other Few-Shot Tasks

Beyond image classification, the proposed method’s core modules exhibit promising applicability to other few-shot learning scenarios, where limited data and feature robustness are critical challenges:

Few-shot object detection: The Adaptive Fusion Convolution Module, designed to capture spatial transformations and deformations, could enhance the localization of novel objects. In scenarios with limited annotated bounding boxes, its ability to adapt to varying object poses and scales (e.g., occluded or rotated objects) may improve the precision of region proposals, complementing existing methods that struggle with geometric variations in rare classes.

Few-shot segmentation: The Median-Enhanced Attention Module, which suppresses background noise while preserving fine-grained details, is well-suited for pixel-level classification of rare classes. For instance, in medical image segmentation (e.g., identifying rare tumors from a handful of annotated slices), the module could focus attention on subtle tissue boundaries obscured by noise, outperforming traditional attention mechanisms that are prone to overfitting to irrelevant background pixels.

Low-resource text classification: The hierarchical metric learning framework, which models feature sets rather than single embeddings, can be adapted to sentence-level representations. By replacing image backbones with pre-trained language models (e.g., BERT), the framework could leverage contextual similarities in low-resource languages, where labeled data for new categories (e.g., domain-specific terminology) is scarce. The median-enhanced attention mechanism could also suppress noisy text patterns (e.g., typos or ambiguous phrases) in underrepresented languages.

To validate these extensions, future work will involve task-specific modifications: adapting the convolution modules to handle spatial coordinates in detection, adjusting attention receptive fields for pixel-level segmentation, and fine-tuning language model outputs for text-based set features. These explorations aim to generalize the method’s noise robustness and multi-scale adaptability to broader few-shot scenarios.

4. Conclusions

This paper proposes a median-enhanced multi-scale adaptive algorithm for few-shot learning, which addresses the limitations of metric learning in capturing global features. The algorithm combines adaptive kernels and variable receptive fields to extract local features, while a median-enhanced attention mechanism refines global features. This dual mechanism enables parallel fusion of local and global features, capturing both detailed and discriminative information. Experimental results on mini-ImageNet, Tiered-ImageNet, and CUB datasets show that our approach outperforms existing models, with significant improvements in classification accuracy. However, the integration of adaptive convolution and attention mechanisms introduces computational complexity. Future work will focus on optimizing the feature extraction process through more efficient convolutions, alternative attention mechanisms, and model compression to reduce computational costs while maintaining high performance.

Author Contributions: C.Y.: methodology, software, formal analysis, investigation, writing—original draft preparation. Z.L.: validation, data curation, investigation. S.Z.: conceptualization, supervision, funding acquisition, writing—review and editing.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61603255.

Data Availability Statement: The datasets used in this study are publicly available.

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

Use of AI and AI-Assisted Technologies: During the preparation of this work, the authors used AI-assisted tools for language polishing. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

References

1. Zhang, X.Y.; Zhou, X.Y.; Lin, M.X.; et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856. https://doi.org/10.1109/CVPR.2018.00716.
2. Wang, Y.X.; Hebert, M. Learning from Small Sample Sets by Combining Unsupervised Meta-Training with CNNs. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
3. Lee, K.; Maji, S.; Ravichandran, A.; et al. Meta-Learning with Differentiable Convex Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10649–10657. https://doi.org/10.1109/CVPR.2019.01091.
4. Afrasiyabi, A.; Lalonde, J.-F.; Gagne, C.; et al. Mixture-Based Feature Space Learning for Few-Shot Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9021–9031. https://doi.org/10.1109/ICCV48922.2021.00891.
5. Wang, C.; Wang, Z.; Liu, W.; et al. A Novel Deep Offline-to-Online Transfer Learning Framework for Pipeline Leakage Detection with Small Samples. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 1–13. https://doi.org/10.1109/TIM.2022.3220302.
6. Rao, S.; Huang, J.; Tang, Z. RdProtoFusion: Refined Discriminative Prototype-Based Multi-Task Fusion for Cross-Domain Few-Shot Learning. *Neurocomputing* **2024**, *599*, 128117. https://doi.org/10.1016/j.neucom.2024.128117.
7. Zhang, C.; Cai, Y.; Lin, G.; et al. DeepEMD: Differentiable Earth Mover’s Distance for Few-Shot Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5632–5648. https://doi.org/10.1109/TPAMI.2022.3217373.
8. Xu, C.; Fu, Y.; Liu, C.; et al. Learning Dynamic Alignment via Meta-Filter for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5178–5187. https://doi.org/10.1109/CVPR46437.2021.00514.
9. Afrasiyabi, A.; Larochelle, H.; Lalonde, J.-F.; et al. Matching Feature Sets for Few-Shot Image Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9004–9014. https://doi.org/10.1109/CVPR52688.2022.00881.
10. Zhang, Y.; Zhou, X.; Wang, N.; et al. DOUN-GNN: Double Nodes Graph Neural Network for Few-Shot Learning. *Neurocomputing* **2025**, *617*, 128970. https://doi.org/10.1016/j.neucom.2024.127625.
11. Chen, H.; Wu, R.; Tao, C.; et al. Multi-Scale Class Attention Network for Diabetes Retinopathy Grading. *Int. J. Network Dyn. Intell.* **2024**, *3*, 100012. https://doi.org/10.32604/ijndi.2023.027425.
12. Ma, C.; Cheng, P.; Cai, C.; et al. Localization and Mapping Method Based on Multimodal Information Fusion and Deep Learning for Dynamic Object Removal. *Int. J. Network Dyn. Intell.* **2024**, *3*, 100008. https://doi.org/10.53941/ijndi.2024.100008.
13. Chen, Z.; Zhang, L.; Tang, J.; et al. Conditional Generative Adversarial Net Based Feature Extraction Along with Scalable Weakly Supervised Clustering for Facial Expression Classification. *Int. J. Network Dyn. Intell.* **2024**, *3*, 100024. https://doi.org/10.53941/ijndi.2024.100024.

14. Li, X.; Li, M.; Yan, P.; et al. Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds. *Int. J. Network Dyn. Intell.* **2023**, *2*, 93–116. https://doi.org/10.53941/ijndi0201006.
15. Rashid, K.I.; Yang, C.; Huang, C.; et al. Fast-DSAGCN: Enhancing Semantic Segmentation with Multifaceted Attention Mechanisms. *Neurocomputing* **2024**, *587*, 127625. https://doi.org/10.1016/j.neucom.2024.127625.
16. Wang, C.; Wang, Z.; Dong, H.; et al. A Novel Prototype-Assisted Contrastive Adversarial Network for Weak-Shot Learning with Applications: Handling Weakly Labeled Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1234–1245. https://doi.org/10.1109/TPAMI.2023.123456.
17. Wang, C.; Wang, Z.; Ma, L.; et al. Subdomain-Alignment Data Augmentation for Pipeline Fault Diagnosis: An Adversarial Self-Attention Network. *IEEE Trans. Ind. Inf.* **2023**, *20*, 1374–1384. https://doi.org/10.1109/TII.2023.3275701.
18. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; et al. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. https://doi.org/10.1109/WACV.2018.00097.
19. Abdelaziz, M.; Zhang, Z. Multi-Scale Kronecker-Product Relation Networks for Few-Shot Learning. *Multimedia Tools Appl.* **2022**, *81*, 6703–6722. https://doi.org/10.1007/s11042-021-11735-w.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929.
21. Vinyals, O.; Blundell, C.; Lillicrap, T.; et al. Matching Networks for One Shot Learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 3637–3645.
22. Lee, S.; Moon, W.; Heo, J.P. Task Discrepancy Maximization for Fine-Grained Few-Shot Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5331–5340. https://doi.org/10.1109/CVPR52688.2022.00526.
23. Chen, Y.; Liu, Z.; Xu, H.; et al. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9042–9051. https://doi.org/10.1109/ICCV48922.2021.00893.
24. Sun, Q.R.; Liu, Y.Y.; Chua, T.S.; et al. Meta-Transfer Learning for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9042–9051. https://doi.org/10.1109/ICCV48922.2021.00893.
25. Simon, C.; Koniusz, P.; Nock, R.; et al. Adaptive Subspaces for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4135–4144. https://doi.org/10.1109/CVPR42600.2020.00419.
26. Snell, J.; Swersky, K.; Zemel, R. Prototypical Networks for Few-Shot Learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 4080–4090.
27. Sung, F.; Yang, Y.; Zhang, L.; et al. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208. https://doi.org/10.1109/CVPR.2018.00131.
28. Huang, H.; Zhang, J.; Zhang, J.; et al. Low-Rank Pairwise Alignment Bilinear Network for Few-Shot Fine-Grained Image Classification. *IEEE Trans. Multimedia* **2021**, *23*, 1666–1680. https://doi.org/10.1109/TMM.2020.3001510.