



Article

Few-Shot Image Classification Algorithm Based on Global–Local Feature Fusion

Lei Zhang ^{1,2}, Xinyu Yang ², Xiyuan Cheng ³, Wenbin Cheng ^{1,2,*}  and Yiting Lin ¹ 

¹ Zhongshan Institute, University of Electronic Science and Technology of China, Zhongshan 528402, China; 202222011529@std.uestc.edu.cn (L.Z.); yitingLin@ieee.org (Y.L.)

² School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; 202322011503@std.uestc.edu.cn

³ School of Automation, Guangdong University of Technology, Guangzhou 510006, China; 2112304365@mail2.gdut.edu.cn

* Correspondence: chengwenbin@zsc.edu.cn

Abstract

Few-shot image classification seeks to recognize novel categories from only a handful of labeled examples, but conventional metric-based methods that rely mainly on global image features often produce unstable prototypes under extreme data scarcity, while local-descriptor approaches can lose context and suffer from inter-class local-pattern overlap. To address these limitations, we propose a Global–Local Feature Fusion network that combines a frozen, pretrained global feature branch with a self-attention based multi-local feature fusion branch. Multiple random crops are encoded by a shared backbone (ResNet-12), projected to Query/Key/Value embeddings, and fused via scaled dot-product self-attention to suppress background noise and highlight discriminative local cues. The fused local representation is concatenated with the global feature to form robust class prototypes used in a prototypical-network style classifier. On four benchmarks, our method achieves strong improvements: Mini-ImageNet 70.31% \pm 0.20 (1-shot)/85.91% \pm 0.13 (5-shot), Tiered-ImageNet 73.37% \pm 0.22/87.62% \pm 0.14, FC-100 47.01% \pm 0.20/64.13% \pm 0.19, and CUB-200-2011 82.80% \pm 0.18/93.19% \pm 0.09, demonstrating consistent gains over competitive baselines. Ablation studies show that (1) naive local averaging improves over global-only baselines, (2) self-attention fusion yields a large additional gain (e.g., +4.50% in 1-shot on Mini-ImageNet), and (3) concatenating global and fused local features gives the best overall performance. These results indicate that explicitly modeling inter-patch relations and fusing multi-granularity cues produces markedly more discriminative prototypes in few-shot regimes.

Keywords: few-shot learning; image classification; metric learning



Academic Editors: Emanuele Frontoni and Arslan Munir

Received: 14 August 2025

Revised: 19 September 2025

Accepted: 30 September 2025

Published: 9 October 2025

Citation: Zhang, L.; Yang, X.; Cheng, X.; Cheng, W.; Lin, Y. Few-Shot Image Classification Algorithm Based on Global–Local Feature Fusion. *AI* **2025**, *6*, 265. <https://doi.org/10.3390/ai6100265>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Deep learning has achieved remarkable success in large-scale vision tasks; however, modern convolutional and transformer backbones typically require substantial amounts of labeled data to establish stable class representations [1–3]. Few-shot image classification, in contrast, seeks to recognize novel categories with only a limited number of labeled samples [4–6]. Metric-based approaches have shown considerable promise in this setting by exploiting the distance or similarity between feature representations. In general, a powerful feature extraction network—either pre-trained on large-scale datasets or strengthened through data augmentation—is employed to obtain high-quality global

features. Nevertheless, when applied to few-shot learning tasks, the scarcity of training samples restricts the model's ability to capture task-relevant information, often leading to suboptimal category representations [7–9]. This challenge becomes even more critical in domains where data security is paramount, such as medical imaging or remote sensing, where the protection of sensitive information imposes additional constraints [10–12]. In contrast, humans exhibit remarkable adaptability in few-shot recognition scenarios. When encountering a novel object, people typically perform a holistic, global inspection followed by focused attention on discriminative local details or surrounding regions. This cognitive process enables humans to extract sufficient task-specific information even from limited observations. As illustrated in Figure 1, when identifying a red-winged blackbird, humans first recognize its global shape and posture as indicative of a bird. Subsequently, distinctive local features—such as the orange-red shoulder patches with yellow margins, along with the beak, tail, and legs—facilitate a more precise classification. This observation suggests that local features can effectively complement global features, thereby producing more robust and informative category prototypes.

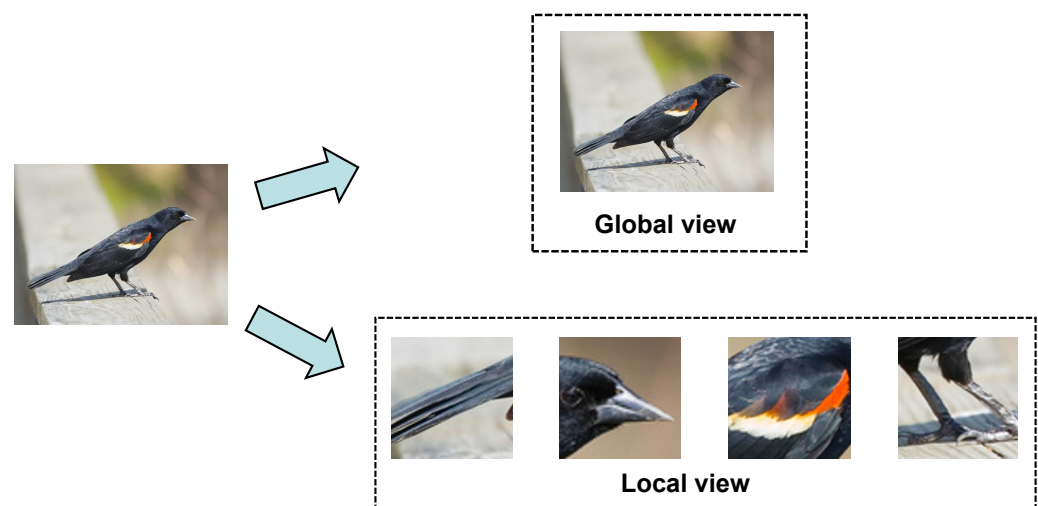


Figure 1. Global and local views of the red-winged blackbird.

Propelled by continuous technological advances, the field of few-shot image classification has attracted considerable scholarly interest, and numerous research findings have emerged [13–16]. In 2023, Ref. [17] proposed a gradual machine-learning (GML) framework for few-shot image classification that departs from the i.i.d. assumption by iteratively labeling target images from easy to hard via factor-graph inference. The method fuses ResNet-12 and WRN-28-10 backbones to extract discriminative features, constructs unary class-centroid and binary k-nearest-neighbor factors, and employs sigmoid influence modeling to propagate knowledge. Extensive experiments on MiniImageNet, TieredImageNet, Cifar-FS, and CUB-200-2011 demonstrate consistent 1–5 % accuracy gains over state-of-the-art inductive and transductive baselines, robust improvements as query-set size increases, and strong cross-domain generalization. These findings advance few-shot learning reliability and open new avenues for gradual inference in computer vision. In 2024, Ref. [18] introduced a novel metric learning method for few-shot image classification combining PatchUp-based feature-space block-level regularization and self-supervised auxiliary loss, integrating feature preprocessing (centering and normalization) for improved generalization capability and classification accuracy, with experiments on datasets like miniImageNet, tieredImageNet, FC-100, and CUB-200-2011 showing significant performance gains (e.g., 2.90% higher 1-shot accuracy on miniImageNet and 3.77% on CUB) and robustness to data scarcity. In 2025, Ref. [19] introduced the MFEHGNN framework that

fuses LiDAR and hyperspectral images through a multimodal feature-enhancement module and a hypergraph neural network, leveraging supervised contrastive learning to capture high-order sample correlations. Experiments across three datasets demonstrated superior few-shot classification accuracy and robustness compared to existing methods, advancing secure and efficient remote-sensing image analysis. Most existing metric-based few-shot classification methods, such as prototypical networks, rely heavily on global features. These methods represent each category by averaging the feature vectors of its support samples extracted by a convolutional neural network (CNN). However, this approach is highly dependent on the capacity of the backbone network and may suffer from large intra-class variation or prototype estimation errors. To address this limitation, several studies have explored local feature-based methods, such as DN4 [20], which utilize deep local descriptors rather than holistic image features. In DN4, local descriptors extracted from convolutional layers are matched via a k-nearest-neighbor classifier to determine the image's class. While this approach captures fine-grained information, it may overlook global context and is susceptible to confusion when local patterns overlap across classes. Consequently, relying solely on local features risks overfitting to irrelevant details and degrading classification performance. Existing metric methods often follow one of two imperfect strategies. The first leverages global (holistic) image features produced by a pre-trained backbone and aggregates support features (e.g., by averaging) into prototypes. While this works well when backbones were trained on large related datasets, it is vulnerable to high intra-class variation and prototype errors when support data are extremely scarce. The second strategy emphasizes local descriptors (e.g., dense deep local features or patch descriptors) which capture fine details useful for fine-grained classes, but local-only methods can ignore global context and are prone to confusion when similar local patterns appear across different classes. Both strategies alone can therefore fail to form prototypes that are both robust and discriminative in ultra-few-shot settings.

Motivated by the human recognition process and the complementary nature of global and local features, we propose a novel few-shot image classification algorithm based on global–local feature fusion. The proposed method incorporates both global representations and informative local cues. Specifically, we extract global features from the original image and generate multiple local patches via random cropping. These patches are independently processed through a shared feature extractor and aggregated using a self-attention mechanism, which learns the relevance of each local feature and assigns adaptive weights accordingly. By emphasizing critical local cues and suppressing less relevant ones, we obtain an enhanced local representation. Finally, the global and aggregated local features are concatenated to form the final class prototype. This fusion strategy enriches the feature space and improves the model's ability to generalize to unseen categories under few-shot settings.

The remaining part of this article is organized as follows. Section 2 describes the overall framework, backbone pre-training and the random-crop local view strategy. Section 3 presents the self-attention based local feature fusion module and the procedure for constructing global–local prototypes. Section 4 reports experimental settings, comparative evaluations, ablation studies (component effectiveness, number of local views, pre-training variants), and visualization analyses. Section 5 concludes the paper and discusses future directions.

2. Overall Framework

The overall framework of the proposed few-shot image classification algorithm based on global–local feature fusion is illustrated in Figure 2. The algorithm mainly consists of a

backbone network pre-trained on a base dataset, a global feature branch, and a local feature branch with multi-local feature fusion based on a self-attention mechanism.

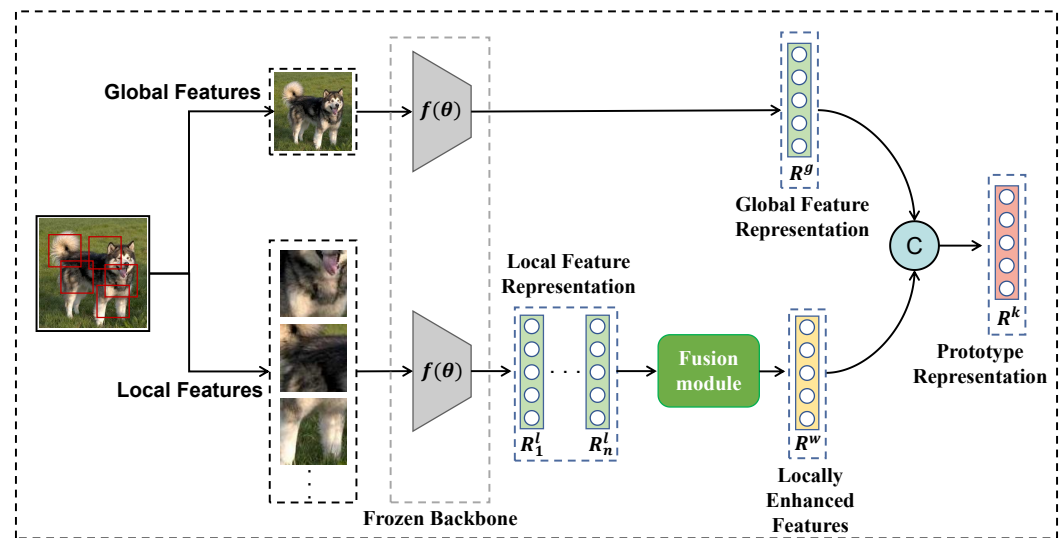


Figure 2. Framework of the proposed few-shot image classification algorithm based on global-local feature fusion.

Assume that we have obtained a powerful feature extractor $f(\theta)$ through meta-learning or pre-training. This extractor is used for few-shot classification tasks. For any training (or test) image sample x , we randomly crop and sample N image patches to represent local views. Let R^g denote the global representation of sample x , and R_l^n denote the n -th local representation, that is,

$$R_g = f(x; \theta) \quad (1)$$

$$R_l^n = f(\tilde{x}_j; \theta) \quad (2)$$

where \tilde{x}_j is the j -th local view obtained by randomly cropping from the original image x .

Through random cropping, multiple local views can be generated from each image sample. These local views are processed by the feature extractor to obtain multiple local feature representations $(R_l^1, R_l^2, \dots, R_l^n)$. However, some of these local views may contain irrelevant information for classification, such as background regions, as shown in Figure 3. Moreover, the feature extraction for each local view is conducted independently, and each feature vector contains information only from a single patch, with inter-patch relationships not being fully exploited.

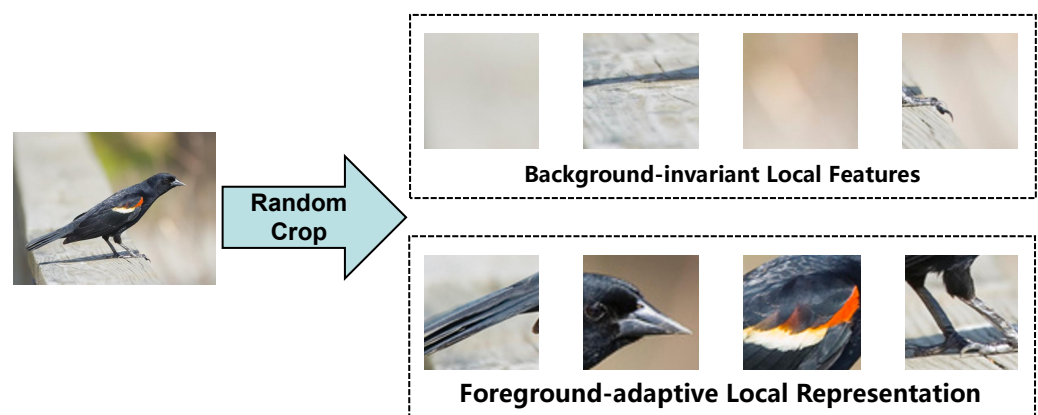


Figure 3. Local views obtained through random cropping.

To effectively utilize the interrelationships among local views and suppress features from irrelevant background regions while enhancing those from the foreground, we feed the local features obtained from the random crops into a self-attention-based fusion module. This module outputs the multi-local fused representation R^w . The detailed structure of this fusion module will be described in Section 3. The final class prototype representation R^k , used for classification, is obtained by concatenating the global feature R^g with the enhanced local feature R^w , which can be expressed as

$$R^k = \text{concat}(R^g, R^w) \quad (3)$$

where concat denotes the concatenation operation.

In each N-way K-shot classification task, for every image in either the support set or the query set, the vector R^k obtained using the above method is used as the image's prototype representation. Following the prototypical network paradigm, we compute the mean feature vector of each class in the support set as its class prototype. The classification of a query image is then performed by computing the Euclidean distances between its feature vector and the class prototypes. The entire classification process is detailed in Algorithm 1. By leveraging both global context and local discriminative details, the proposed fusion strategy significantly enhances few-shot image classification performance.

Algorithm 1 Few-shot classification via global–local feature fusion

1: **Input:**

- Support set $S = \{(x_i, y_i)\}_{i=1}^{N \times K}$
- Query image x_q
- Pre-trained feature extractor f_θ
- Number of local views N

2: **Output:** Predicted label of the query image \hat{y}_q

▷ Prototype computation

3: Initialize prototype set $\mathcal{P} \leftarrow \emptyset$

4: **for** each class $c \in \{1, \dots, N\}$ **do**

5: Initialize feature list $\mathcal{F}_c \leftarrow \emptyset$

6: **for** each sample $(x_i, y_i) \in S$ where $y_i = c$ **do**

7: Extract global feature $R_g \leftarrow f_\theta(x_i)$

8: Generate local views $\{\tilde{x}_i^1, \dots, \tilde{x}_i^N\} \leftarrow \text{RandomCrop}(x_i, N)$

9: Extract local features $[R_i^1, \dots, R_i^N] \leftarrow [f_\theta(\tilde{x}_i^1), \dots, f_\theta(\tilde{x}_i^N)]$

10: Fuse features $R_w \leftarrow \text{SelfAttention}([R_i^1, \dots, R_i^N])$

▷ See Section 3

11: Concatenate features $R^k \leftarrow \text{concat}(R_g, R_w)$

12: $\mathcal{F}_c \leftarrow \mathcal{F}_c \cup \{R^k\}$

13: **end for**

14: Compute prototype $p_c \leftarrow \frac{1}{|\mathcal{F}_c|} \sum_{R \in \mathcal{F}_c} R$

15: $\mathcal{P} \leftarrow \mathcal{P} \cup \{p_c\}$

16: **end for**

▷ Query sample classification

17: Generate local crops $\{\tilde{x}_q^1, \dots, \tilde{x}_q^N\} \leftarrow \text{RandomCrop}(x_q, N)$

18: Extract query feature:

19: $R_q^k \leftarrow \text{concat}(f_\theta(x_q), \text{SelfAttention}([f_\theta(\tilde{x}_q^1), \dots, f_\theta(\tilde{x}_q^N)]))$

20: **for** each class $c \in \{1, \dots, N\}$ **do**

21: Compute distance $d_c \leftarrow \|R_q^k - p_c\|_2^2$

22: **end for**

23: Determine label $\hat{y}_q \leftarrow \arg \min_c (d_1, \dots, d_N)$

24: **return** \hat{y}_q

3. Feature Fusion Module

3.1. Attention Mechanism

The design of the attention mechanism is inspired by insights from human cognitive processes. In cognitive science, it is well established that humans possess a limited capacity for information processing, which compels them to selectively focus on salient parts of the input while disregarding less relevant details. This selective attention serves as an effective strategy for allocating limited visual processing resources. For example, in the image recognition task illustrated in Figure 4, observers typically concentrate on the animal's face while paying little attention to the background. The corresponding attention map highlights darker regions where attention is concentrated, thereby indicating that the animal is most likely a lion.

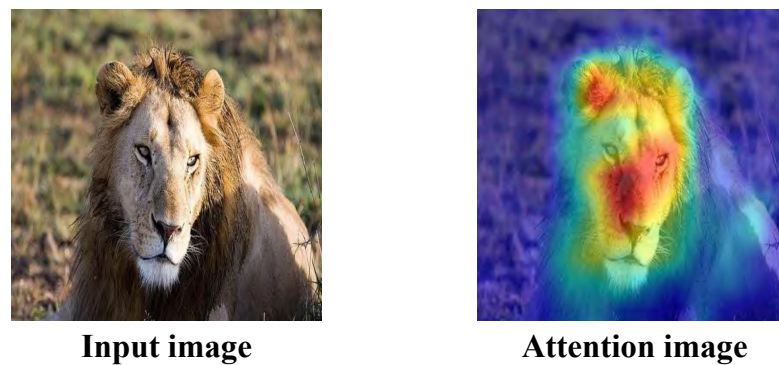


Figure 4. Original image and corresponding attention map.

In deep learning, the attention mechanism has been introduced to mitigate information overload and to allocate computational resources more effectively to critical regions under limited capacity. In general, increasing the number of network parameters enhances representational power and allows the encoding of richer information. However, this also raises the risk of redundancy and information overload. The attention mechanism addresses this issue by directing the model to focus on the most relevant parts of the input while suppressing or filtering out irrelevant details, thereby improving both efficiency and task accuracy. Through this process, the model can rapidly extract high-value information from large volumes of data using limited attentional resources.

Typically, the attention mechanism comprises three key components: Query, Key, and Value. The Query acts as an internal cue, representing the feature vector that guides the model to retrieve relevant information. The Key is the external reference, representing salient features of input elements. The Value contains the actual information content and is paired with the Key. Attention is computed by measuring the similarity between the Query and each Key, assigning weights to the Values accordingly to produce the final output. The computation process is illustrated in Figure 5. Given a query vector \mathbf{Q} , the model computes the similarity score \mathbf{S}_i between \mathbf{Q} and each key \mathbf{K}_i as follows:

$$\mathbf{S}_i = \mathbf{Q} \cdot \mathbf{K}_i \quad (4)$$

Then, the similarity scores are normalized using the softmax function to produce attention weights α_i :

$$\alpha_i = \frac{\exp(\mathbf{S}_i)}{\sum_{j=1}^n \exp(\mathbf{S}_j)} \quad (5)$$

Finally, the attention output is obtained by the weighted sum of the corresponding values:

$$\text{Attention Value} = \sum_{i=1}^n \alpha_i V_i \quad (6)$$

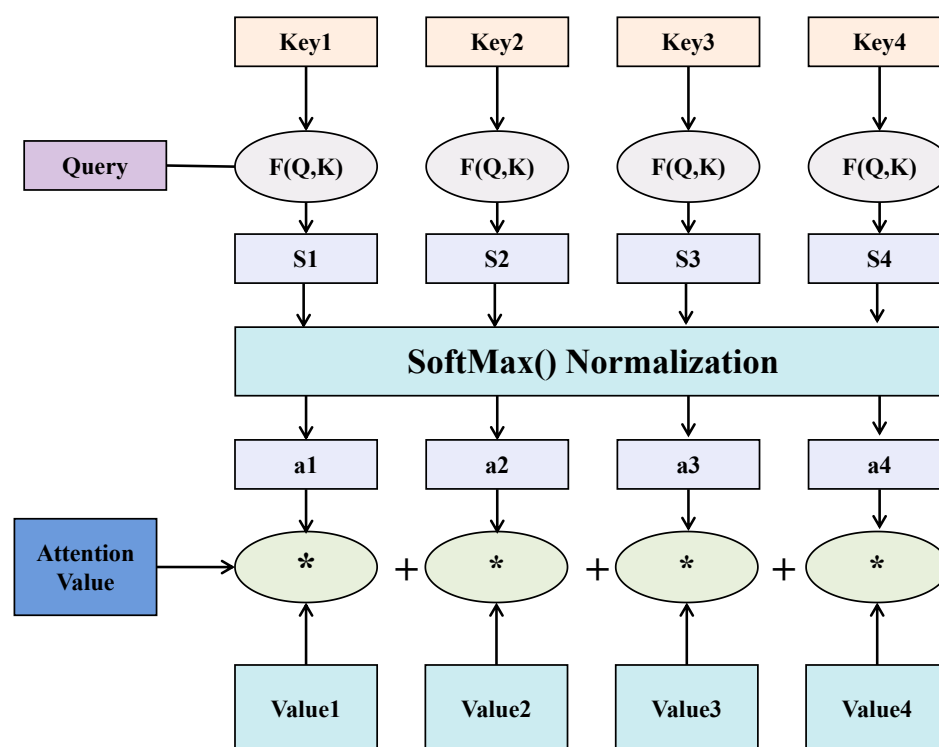


Figure 5. Computation process of the attention mechanism.

3.2. Self-Attention Mechanism

Self-attention, as a pivotal variant of the attention mechanism, is designed to address the limitations of traditional neural networks in processing multi-input data. In natural language processing (NLP) tasks such as machine translation, part-of-speech tagging, and semantic analysis, the inputs typically consist of variable-length sequences with complex interdependencies. Conventional fully connected networks struggle to capture these dependencies effectively, often resulting in suboptimal performance. Self-attention overcomes this limitation by dynamically modeling the interactions among different parts of the input through an internal attention mechanism, thereby enhancing the model's ability to extract task-relevant information.

Unlike conventional attention mechanisms that rely on external signals, self-attention exploits the internal correlations within a single input, where the Query, Key, and Value vectors are all derived from the same source. This design allows the model to emphasize informative components while suppressing irrelevant ones. As a result, self-attention not only captures long-range dependencies more effectively but also enables parallel computation, significantly improving training efficiency. Furthermore, by dynamically assigning attention weights, it filters out noise and reduces redundancy, thereby enhancing performance across a wide range of tasks. A milestone in this direction was the introduction of the Transformer architecture by Vaswani et al. in 2017 through the seminal work *Attention Is All You Need*. Unlike RNN- or CNN-based architectures, the Transformer relies entirely on self-attention, which facilitates full parallelism and effective modeling of long-range dependencies, thus achieving substantial efficiency gains in handling long sequences.

3.3. Self-Attention-Based Local Feature Fusion

Inspired by the self-attention mechanism and the Transformer architecture, we propose a self-attention-based local feature fusion module. The overall structure is illustrated in Figure 6. The detailed procedure is shown in the following.

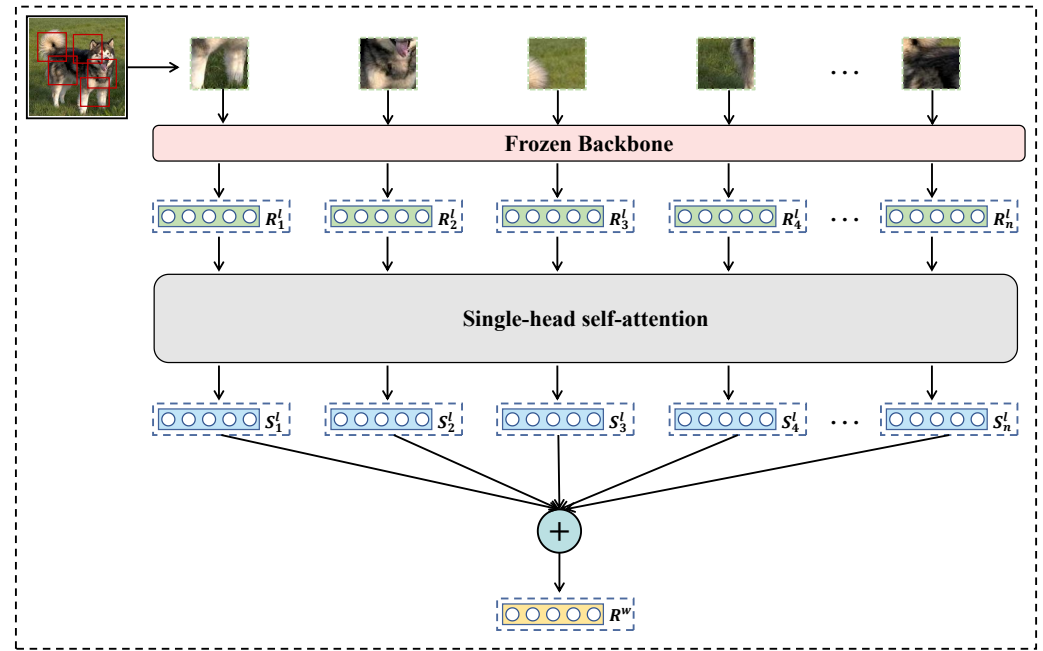


Figure 6. Diagram of the self-attention-based local feature fusion module.

- (1) **Image Preprocessing:** Multiple local views of the input image are generated by applying a RandomResizedCrop operation, producing patches of size 84×84 from different regions of the original image.
- (2) **Feature Extraction:** A pre-trained feature extractor $f(\theta)$ is used to encode each cropped patch, resulting in a set of local feature vectors $(R_1^l, R_2^l, \dots, R_n^l)$.
- (3) **Feature Fusion:** The extracted local features are then fused using a self-attention mechanism. Specifically, each local feature R_i^l is projected through three separate fully connected layers W^Q , W^K , and W^V to generate its corresponding Query, Key, and Value vectors:

$$Q_i = R_i^l W^Q, \quad K_i = R_i^l W^K, \quad V_i = R_i^l W^V \quad (7)$$

For each query Q_i , the similarity with all keys K_j is computed to obtain attention scores A_{ij} using the scaled dot-product attention:

$$A_{ij} = \frac{Q_i K_j^T}{\sqrt{d_k}} \quad (8)$$

where d_k is the dimensionality of the key vectors, used to stabilize gradients.

Next, a softmax function is applied to normalize the attention scores, and the weighted sum of the values V_j is computed to yield the fused local feature:

$$R_i^l = \sum_{j=1}^n \frac{\exp(A_{ij})}{\sum_{k=1}^n \exp(A_{ik})} V_j \quad (9)$$

In this way, each local feature R_i^l incorporates contextual information from all other patches, enhancing its global representation.

Finally, the enhanced local features ($R_1^l, R_2^l, \dots, R_n^l$) are aggregated by summation to produce the final fused representation R^w . This module effectively integrates local features and global context, thereby enhancing the model's expressive power and overall performance.

4. Experimental Results and Analysis

4.1. Experimental Setup

All experiments in this study were conducted under a unified experimental environment. The code was implemented in Python, and the networks were built using the PyTorch framework. An RTX 2080 Ti (11GB) GPU was used for training. The main software and hardware configurations used in the experiments are shown in Table 1.

Table 1. Experimental environment.

Device	Configuration
Programming Language	Python 3.8
Framework	PyTorch 1.11.0
Operating System	Ubuntu 20.04
Development Environment	PyCharm 2022
CUDA Version	CUDA 11.3
GPU	RTX 2080 Ti (11GB) \times 1
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8255C @ 2.50GHz

4.2. Training Parameters

The training process of our model comprises two stages: pre-training and adaptation. In the pre-training stage, to ensure a fair comparison with state-of-the-art methods, we adopt ResNet-12—widely used in few-shot image classification—as the backbone network. By default, the backbone is pre-trained following the two-stage strategy described in Section 3, which integrates Patchup regularization with a self-supervised auxiliary loss. For all datasets, the number of local views per image is fixed at 10.

In the adaptation stage, the pre-trained backbone weights are used for initialization, and its parameters are frozen. The remaining modules are then trained in an end-to-end manner using stochastic gradient descent (SGD) with an initial learning rate of 0.0001, momentum of 0.9, and weight decay of 5×10^{-4} . The batch size is set to 64, and training is performed for 50 epochs.

For validation, classification is conducted on the validation set in few-shot learning mode. Specifically, 500 episodes are randomly sampled under the 5-way K -shot setting ($K = 1$ or 5). In each episode, K samples per class are selected as the support set, and 15 samples per class serve as the query set.

In the testing phase, performance is evaluated on 10,000 episodes. Accuracy, reported with a 95% confidence interval, is used as the evaluation metric to provide an objective and comprehensive assessment of the model.

4.3. Comparative Results and Analysis

To validate the effectiveness of the proposed method, this section presents a comparative analysis against classical and state-of-the-art few-shot image classification methods. These include Prototypical Networks (ProtoNet), Matching Networks (MatchingNet), Task-Adaptive Metric Learning (TADAM), and Meta-Baseline. The results are summarized in Tables 2 and 3. **Bolded values** indicate the best performance in each column, underlined values indicate the second-best, and “—” indicates unreported results in the original literature.

Table 2. Comparative experimental results on Mini-ImageNet and FC-100 datasets.

Method	Backbone	Mini-ImageNet		FC-100	
		1-Shot	5-Shot	1-Shot	5-Shot
MatchingNet [21]	ResNet12	63.08 \pm 0.80	75.99 \pm 0.60	-	-
ProtoNet [22]	ResNet12	60.37 \pm 0.83	78.02 \pm 0.57	37.53 \pm 0.40	38.39 \pm 0.40
TADAM [23]	ResNet12	58.50 \pm 0.30	76.70 \pm 0.30	40.10 \pm 0.40	56.10 \pm 0.40
Meta-Baseline [24]	ResNet12	63.17 \pm 0.23	79.26 \pm 0.17	-	-
DeepEMD [25]	ResNet12	65.91 \pm 0.82	82.41 \pm 0.56	46.60 \pm 0.26	63.22 \pm 0.71
MACF [26]	ResNet12	67.30 \pm 0.23	83.31 \pm 0.17	46.36 \pm 0.40	62.99 \pm 0.41
Lsfl [27]	ResNet12	64.67 \pm 0.49	81.79 \pm 0.18	43.60 \pm 0.11	60.12 \pm 0.17
SSFormers [28]	ResNet12	67.25 \pm 0.24	82.75 \pm 0.20	43.72 \pm 0.21	58.92 \pm 0.18
QSFormer [29]	ResNet12	65.24 \pm 0.28	79.96 \pm 0.20	46.51 \pm 0.26	61.58 \pm 0.25
MetaOptNet [30]	ResNet12	64.09 \pm 0.62	80.00 \pm 0.45	47.20 \pm 0.60	62.50 \pm 0.60
S2M2R [31]	ResNet18+	64.06 \pm 0.18	80.58 \pm 0.12	-	-
LEO [32]	WRN-28-10+	61.76 \pm 0.08	77.59 \pm 0.12	-	-
MATANet [33]	WRN-28-10+	62.43 \pm 0.23	79.02 \pm 0.72	-	-
Our	ResNet12	70.31 \pm 0.20	85.91 \pm 0.13	47.01 \pm 0.20	64.13 \pm 0.19

Table 3. Comparative experimental results on Tiered-ImageNet and CUB datasets.

Method	Backbone	Tiered-ImageNet		CUB-200-2011	
		1-Shot	5-Shot	1-Shot	5-Shot
MatchingNet [21]	ResNet12	68.50 \pm 0.92	80.60 \pm 0.71	71.87 \pm 0.85	85.08 \pm 0.57
ProtoNet [22]	ResNet12	65.65 \pm 0.92	83.40 \pm 0.65	66.09 \pm 0.92	82.50 \pm 0.58
Meta-Baseline [24]	ResNet12	68.62 \pm 0.27	83.74 \pm 0.18	-	-
DeepEMD [25]	ResNet12	71.16 \pm 0.87	86.03 \pm 0.58	75.65 \pm 0.83	88.69 \pm 0.50
MACF [26]	ResNet12	71.92 \pm 0.52	85.89 \pm 0.17	81.04 \pm 0.43	91.53 \pm 0.17
Lsfl [27]	ResNet12	71.17 \pm 0.52	86.23 \pm 0.22	-	-
GLoFA [34]	ResNet12	69.75 \pm 0.33	83.58 \pm 0.42	-	-
QSFormer [29]	ResNet12	72.47 \pm 0.31	85.43 \pm 0.22	75.44 \pm 0.29	86.30 \pm 0.19
AFHN [35]	ResNet18+	-	-	70.53 \pm 1.01	83.95 \pm 0.63
S2M2R [31]	ResNet18+	-	-	71.43 \pm 0.43	85.55 \pm 0.52
LEO [32]	WRN-28-10+	66.33 \pm 0.05	81.44 \pm 0.09	-	-
Our	ResNet12	73.37 \pm 0.22	87.62 \pm 0.14	82.80 \pm 0.18	93.19 \pm 0.09

On the Mini-ImageNet dataset, our method achieves an accuracy of 70.31% in the 1-shot task, representing a notable improvement of 3.01% over the state-of-the-art MACF. In the 5-shot task, our approach reaches 85.91%, outperforming MACF by 2.60%. On the FC-100 dataset, our method also performs competitively, achieving 47.01% in the 1-shot setting (a 0.19% improvement over MetaOptNet) and 64.13% in the 5-shot setting (a 0.91% improvement over DeepEMD).

On the Tiered-ImageNet dataset, our method obtains 73.37% accuracy in the 1-shot setting—1.45% higher than MACF—and 87.62% in the 5-shot setting, surpassing MACF by 1.73%. These results demonstrate the robustness of the proposed approach on large-scale benchmarks.

On the fine-grained dataset CUB-200-2011, our method is particularly effective. It achieves 82.80% accuracy in the 1-shot task, surpassing MACF by 1.76%, and 93.19% in the 5-shot task, improving upon MACF by 1.66%. The performance gain can be attributed to the sensitivity of fine-grained classification to subtle visual cues. By enhancing the extraction and alignment of discriminative local features—such as beak shape and feather texture—through local feature fusion, our method ensures that these critical details dominate classification outcomes.

Furthermore, we compare our approach with MACF, which leverages multi-granularity cross-layer fusion, and MATANet, which applies adaptive fusion of multi-scale features. Across all datasets, our method consistently outperforms these approaches, further validating the effectiveness of the proposed fusion strategy.

In summary, the experimental results across multiple datasets confirm that our method substantially advances few-shot image classification performance. The findings highlight the strong generalization capability of the proposed global–local feature fusion strategy. By combining holistic global context with fine-grained local details, our model effectively captures key discriminative features under limited data conditions, thereby enabling more accurate classification. The advantage is particularly pronounced in fine-grained scenarios, underscoring the robustness and effectiveness of the proposed method in complex classification tasks.

4.4. Ablation Study Results and Analysis

4.4.1. Component Effectiveness Analysis

To systematically evaluate the contribution of each key component in our proposed method, we conduct a series of ablation studies on the Mini-ImageNet dataset. The experimental procedure is as follows: first, the backbone network is pre-trained using our proposed two-stage pre-training algorithm. Then, we compare four variants to investigate the individual roles of each component:

- (1) **Global Features:** Only the global branch is utilized, where the original image is directly fed into the backbone network to extract global features. These features are used as the class prototypes for classification.
- (2) **Local Features:** Multiple local views are obtained via random cropping and input into the backbone for feature extraction. However, instead of using our designed feature fusion module, the local feature vectors are simply averaged to form the class prototype.
- (3) **Local Feature Fusion:** After extracting features from randomly cropped local views, the proposed feature fusion module is applied to obtain a fused feature vector, which is then used as the class prototype for classification.
- (4) **Global–Local Feature Fusion:** The global features are concatenated with the enhanced local features obtained via the fusion module to form the final class prototype. This configuration represents our proposed method.

In this study, the number of local views is set to 10. The ablation results are summarized in Table 4.

Table 4. Ablation results of different components on Mini-ImageNet.

Method	5-Way 1-Shot	5-Way 5-Shot
Global Features	63.73 ± 0.20	83.92 ± 0.13
Local Features	65.26 ± 0.20	84.92 ± 0.13
Local Feature Fusion	69.76 ± 0.20	85.42 ± 0.13
Global–Local Feature Fusion (ours)	70.31 ± 0.20	85.91 ± 0.12

The results indicate a progressive performance gain as more components are introduced. Using only global features yields a baseline performance of 63.73% and 83.92% for the 1-shot and 5-shot tasks, respectively, validating the backbone’s representational capacity. Introducing local features obtained via random cropping and averaging improves classification accuracy to 65.26% (1-shot) and 84.92% (5-shot), demonstrating that local features contribute complementary information. However, naive averaging may fail to capture interactions among local features.

By incorporating the proposed feature fusion module, more discriminative prototypes are formed. Compared to simple averaging, the fusion module improves performance by 4.50% (1-shot) and 0.50% (5-shot), confirming its effectiveness in deeply integrating local information.

Finally, concatenating global and fused local features leads to the best performance: 70.31% (1-shot) and 85.91% (5-shot). This suggests that global features offer contextual information while local features provide fine-grained details. Their synergy significantly enhances classification accuracy.

It is worth noting that the performance gain from the fusion module is larger in the 1-shot setting (+4.50%) than in the 5-shot setting (+0.50%). This is likely due to the model's greater reliance on fine-grained local modeling when sample size is extremely limited. As support samples increase, global statistical features become more stable, and the marginal benefit of optimizing local features decreases. The global–local fusion strategy effectively balances representation robustness and discriminability, yielding optimal performance under varying sample sizes.

In summary, the ablation results validate the rationality of our component designs: local views enhance fine-grained expressiveness, the fusion module effectively integrates local semantics, and the global–local fusion leverages multi-granularity features for improved generalization.

4.4.2. Impact of the Number of Local Views

To systematically assess how the number of randomly cropped local views affects performance, we conduct ablation experiments on the Mini-ImageNet dataset under both 5-way 1-shot and 5-way 5-shot settings. We vary the number of local views (2, 4, 6, 8, 10, 12, 14, 16, 18, 20) and record the classification accuracy of three variants: local features, local feature fusion, and global–local feature fusion.

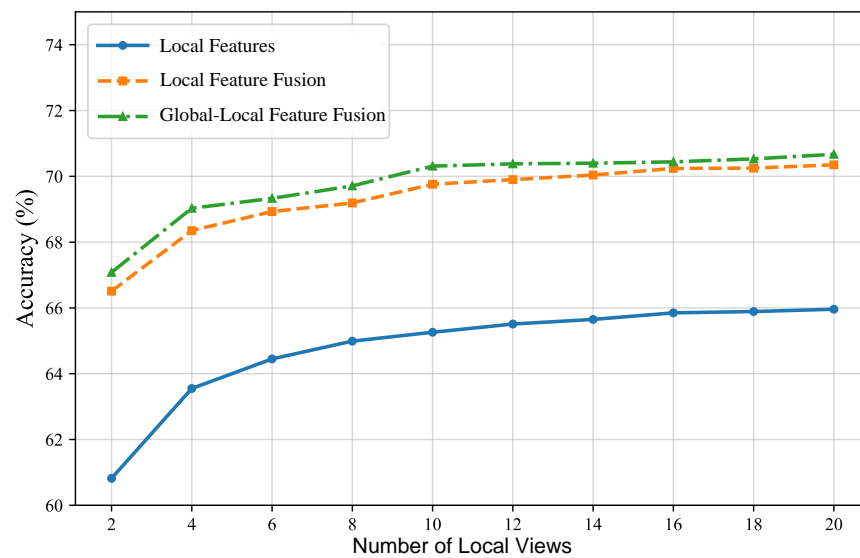
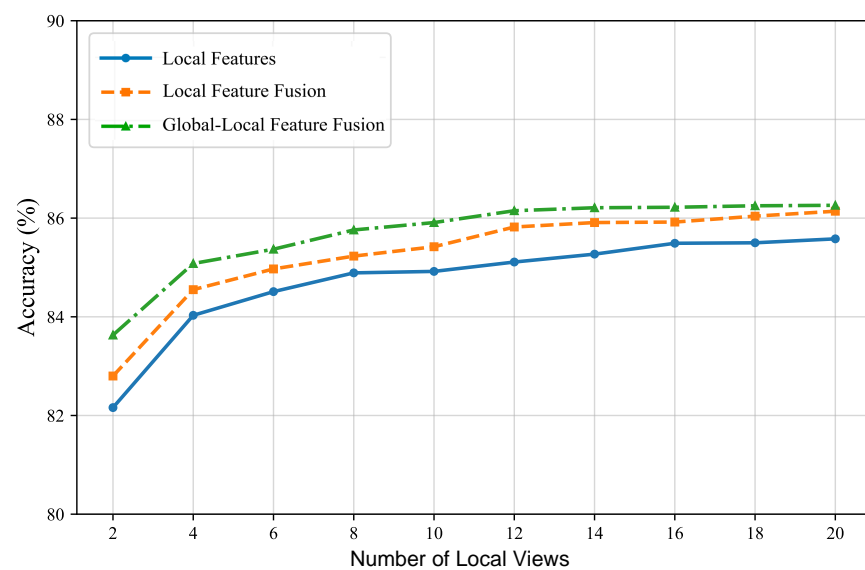
Based on the experimental results presented in Tables 5 and 6, this section analyzes the impact of the number of local views from three perspectives: the performance trend versus local view count, the effectiveness of feature fusion strategies, and the sensitivity disparity across task scenarios. To visually present the experimental outcomes, the classification accuracies of ablation studies on the Mini-ImageNet dataset are visualized in Figures 7 and 8. These figures clearly illustrate the variation trends of classification accuracy under different local view counts, providing intuitive evidence for subsequent analysis.

Table 5. Impact of local view count on 5-way 1-shot accuracy.

No. of Local Views	Method		
	Local Features	Local Feature Fusion	Global–Local Fusion
2	60.82 ± 0.21	66.51 ± 0.20	67.08 ± 0.20
4	63.55 ± 0.21	68.35 ± 0.20	69.03 ± 0.20
6	64.45 ± 0.21	68.93 ± 0.20	69.33 ± 0.20
8	64.99 ± 0.21	69.19 ± 0.20	69.71 ± 0.20
10	65.26 ± 0.21	69.76 ± 0.20	70.31 ± 0.20
12	65.51 ± 0.21	69.90 ± 0.20	70.38 ± 0.20
14	65.65 ± 0.21	70.04 ± 0.20	70.40 ± 0.20
16	65.85 ± 0.21	70.24 ± 0.20	70.44 ± 0.20
18	65.89 ± 0.21	70.25 ± 0.20	70.53 ± 0.20
20	65.96 ± 0.21	70.35 ± 0.20	70.67 ± 0.20

Table 6. Impact of local view count on 5-way 5-shot accuracy.

No. of Local Views	Method		
	Local Features	Local Feature Fusion	Global–Local Fusion
2	82.16 \pm 0.14	82.80 \pm 0.14	83.63 \pm 0.13
4	84.03 \pm 0.13	84.55 \pm 0.13	85.08 \pm 0.12
6	84.51 \pm 0.14	84.97 \pm 0.13	85.37 \pm 0.12
8	84.89 \pm 0.13	85.23 \pm 0.13	85.76 \pm 0.13
10	84.92 \pm 0.13	85.42 \pm 0.13	85.91 \pm 0.13
12	85.11 \pm 0.13	85.82 \pm 0.12	86.15 \pm 0.12
14	85.27 \pm 0.13	85.91 \pm 0.13	86.21 \pm 0.12
16	85.49 \pm 0.13	85.92 \pm 0.13	86.22 \pm 0.12
18	85.50 \pm 0.13	86.04 \pm 0.13	86.25 \pm 0.12
20	85.58 \pm 0.13	86.14 \pm 0.12	86.26 \pm 0.12

**Figure 7.** Comparison of experimental results under varying numbers of local views in the 5-way 1-shot setting.**Figure 8.** Comparison of experimental results under varying numbers of local views in the 5-way 5-shot setting.

Performance Trend versus Local View Count: Under both 1-shot and 5-shot scenarios, all three methodologies exhibit monotonically increasing classification accuracy with additional local views. For example, in the 1-shot task, the global–local fusion method improves from 67.08% (2 views) to 70.67% (20 views), marking a 3.59 percentage point increase. In the 5-shot task, it rises from 83.63% to 86.26%, a gain of 2.63 percentage points. However, beyond 14 views, performance improvements diminish significantly (e.g., only 0.27% increase from 14 to 20 views in 1-shot). This saturation effect suggests increased feature redundancy among local views, limiting information gain from additional views.

Effectiveness of Feature Fusion Strategies: Feature fusion strategies significantly enhance performance at identical view counts. With 10 views in the 1-shot task, the local-only method achieves 65.26% accuracy, while local feature fusion reaches 69.76%—a 4.50 percentage point difference. The global–local fusion method further improves accuracy to 70.31%, gaining an additional 0.55 percentage points. This demonstrates that feature fusion effectively integrates multi-view information to mitigate overfitting in few-shot settings. Global features additionally provide spatial context priors that complement local features, with the global–local fusion showing greater advantages under limited views.

Sensitivity Disparity Across Task Scenarios: The 1-shot task exhibits higher sensitivity to local view counts than the 5-shot task. When views increase from 2 to 20, the performance gain in 1-shot (3.59%) is $1.37\times$ higher than in 5-shot (2.63%), indicating that view diversity is more critical for enhancing generalization under extreme data scarcity. Moreover, performance variations among the three methods are smaller in 5-shot tasks (max gap: 0.68% at 20 views) versus 1-shot tasks (max gap: 4.71%), further validating the necessity of fusion strategies in ultra-few-shot conditions.

In summary, increasing local view counts enhances few-shot classification performance but with marginally diminishing returns. Feature fusion strategies, particularly global–local fusion, substantially improve view utilization efficiency, with more pronounced benefits in 1-shot scenarios. Practical applications should balance the number of randomly cropped local views according to computational resources and task requirements.

4.4.3. Model Effectiveness Under Different Pre-Training Strategies

The training pipeline of our model consists of two stages: pre-training and adaptation. The pre-training stage focuses on training the backbone network, while the adaptation stage initializes the backbone with pre-trained weights and freezes its parameters. The remaining modules are then trained in an end-to-end manner for task-specific adaptation. To further investigate the effectiveness of our proposed global–local feature fusion strategy under different pre-training paradigms, we conduct ablation experiments on the Mini-ImageNet dataset. Three pre-training strategies are compared:

- (1) **Meta-learning:** This strategy adopts an episode-based training paradigm, where 100 meta-tasks are sampled per epoch and the model is trained for 500 epochs.
- (2) **Standard Supervised Learning:** This strategy follows a conventional supervised training pipeline. A prototype classifier is used for few-shot validation, with a total of 100 training epochs.
- (3) **Two-Stage Pre-training:** This strategy combines PatchUp regularization with a self-supervised rotation prediction task for joint optimization.

We record the classification accuracy of the global feature, local feature fusion, and global–local feature fusion variants under each pre-training setting. All pre-training strategies share the same initial learning rate (0.1), cosine learning rate decay, and batch size (128). To avoid interference from data augmentation, all methods use the same standard preprocessing pipeline, including random horizontal flipping and color jittering. The adaptation stage follows the same configuration as described in Section 4.2. For all experiments,

ResNet12 is used as the backbone network, and the number of local views is fixed at 10. Performance is evaluated on both 5-way 1-shot and 5-way 5-shot classification tasks. Results are reported in Tables 7 and 8.

Table 7. Performance comparison under different pre-training strategies on 5-way 1-shot tasks.

Meta-Learning	Supervised Learning	Two-Stage Pre-Training	Method		
			Global Features	Local Feature Fusion	Global–Local Fusion
✓	✓	✓	50.22 ± 0.20	56.01 ± 0.20	56.55 ± 0.20
			58.30 ± 0.20	65.48 ± 0.20	66.30 ± 0.20
			63.73 ± 0.20	69.76 ± 0.20	70.31 ± 0.20

Table 8. Performance comparison under different pre-training strategies on 5-way 5-shot tasks.

Meta-Learning	Supervised Learning	Two-Stage Pre-Training	Method		
			Global Features	Local Feature Fusion	Global–Local Fusion
✓	✓	✓	69.07 ± 0.17	71.62 ± 0.17	72.04 ± 0.17
			80.14 ± 0.14	82.06 ± 0.14	82.64 ± 0.13
			83.92 ± 0.13	85.42 ± 0.13	85.91 ± 0.12

The results demonstrate the general applicability and superiority of our proposed global–local feature fusion strategy across all three pre-training regimes. As shown in Tables 7 and 8, global–local fusion consistently outperforms single-feature baselines.

For instance, under the standard supervised learning setting, global–local fusion improves over global features by 8.00% and over local feature fusion by 0.82% in the 1-shot task. Under the meta-learning setting, it surpasses global features by 6.33% and local feature fusion by 0.54%. These results highlight the complementary nature of global semantic context and local fine-grained details, which together overcome the limitations of feature expressiveness inherent to different pre-training approaches. The fusion mechanism enhances the model’s capacity to capture discriminative cues, thereby yielding superior performance in few-shot scenarios.

4.5. Visualization Analysis

To comprehensively evaluate the effectiveness of the proposed method, we perform an in-depth visualization analysis of the generated features using the t-SNE technique. Specifically, we visualize features extracted by two methods: one using only global features, and the other employing global–local feature fusion (with the number of local views set to 10). In the experiments, we utilize models trained under the 5-way 1-shot setting on both the natural image dataset Mini-ImageNet and the fine-grained dataset CUB-200-2011. From the test sets of these two datasets, we randomly select five different classes and choose 50 images from each class for feature extraction and visualization. The final visualization results are shown in Figures 9 and 10.

The visualization results on both datasets reveal that when only global features are used, the feature distributions in the t-SNE space exhibit clear dispersion. Specifically, there is notable boundary ambiguity and overlap among different class features, indicating that relying solely on global representations is insufficient. The limited discriminative capacity may hinder the classifier from effectively distinguishing between similar classes. This limitation likely stems from global features’ over-reliance on coarse semantic information, while neglecting local discriminative cues crucial for fine-grained categorization.

In contrast, with the proposed global–local feature fusion strategy, the structural organization of the feature space improves significantly. The fused features form compact intra-class clusters and show increased inter-class separability in the t-SNE space. This improvement is mainly attributed to the effective complementarity of multi-scale features:

global features provide contextual semantic understanding, while local features focus on fine-grained, discriminative details. Their synergy enhances both intra-class consistency and inter-class distinctiveness.

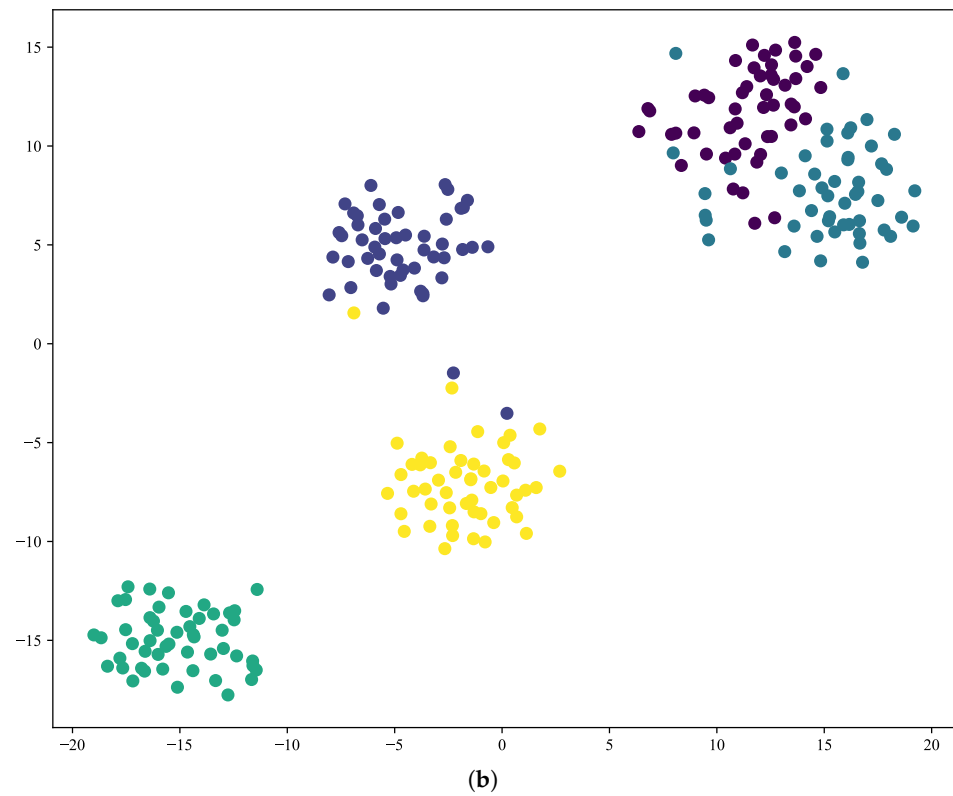
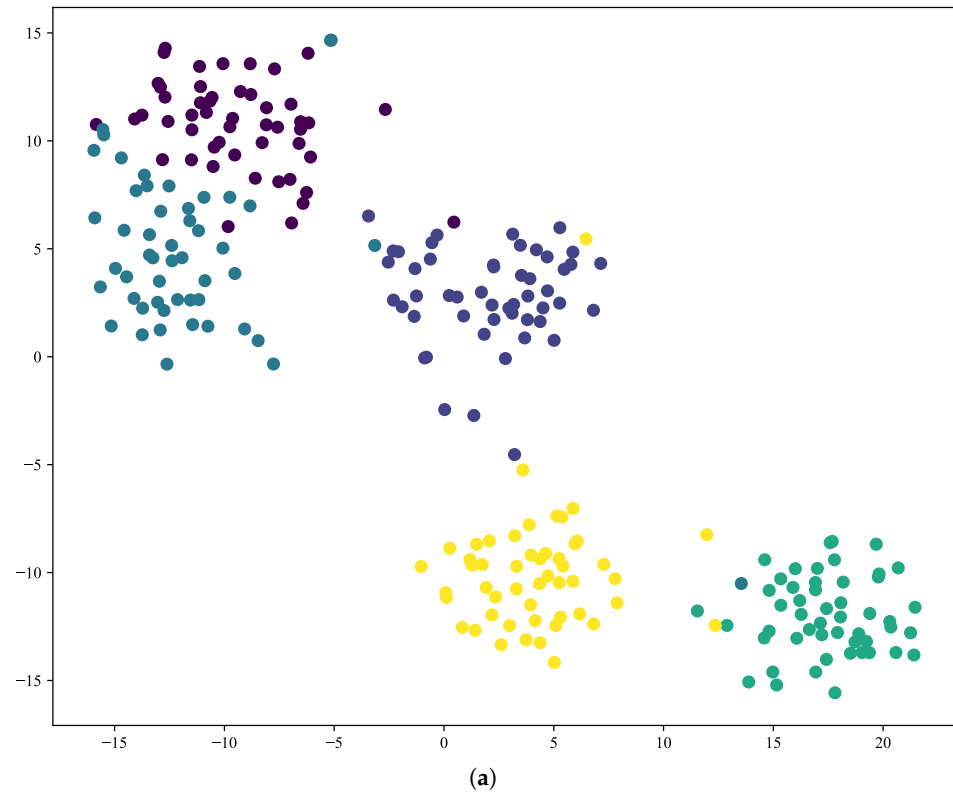


Figure 9. Visualization of feature distribution of global features and global–local feature fusion on Mini-ImageNet: (a) using only global features; (b) global–local feature fusion (Ours).

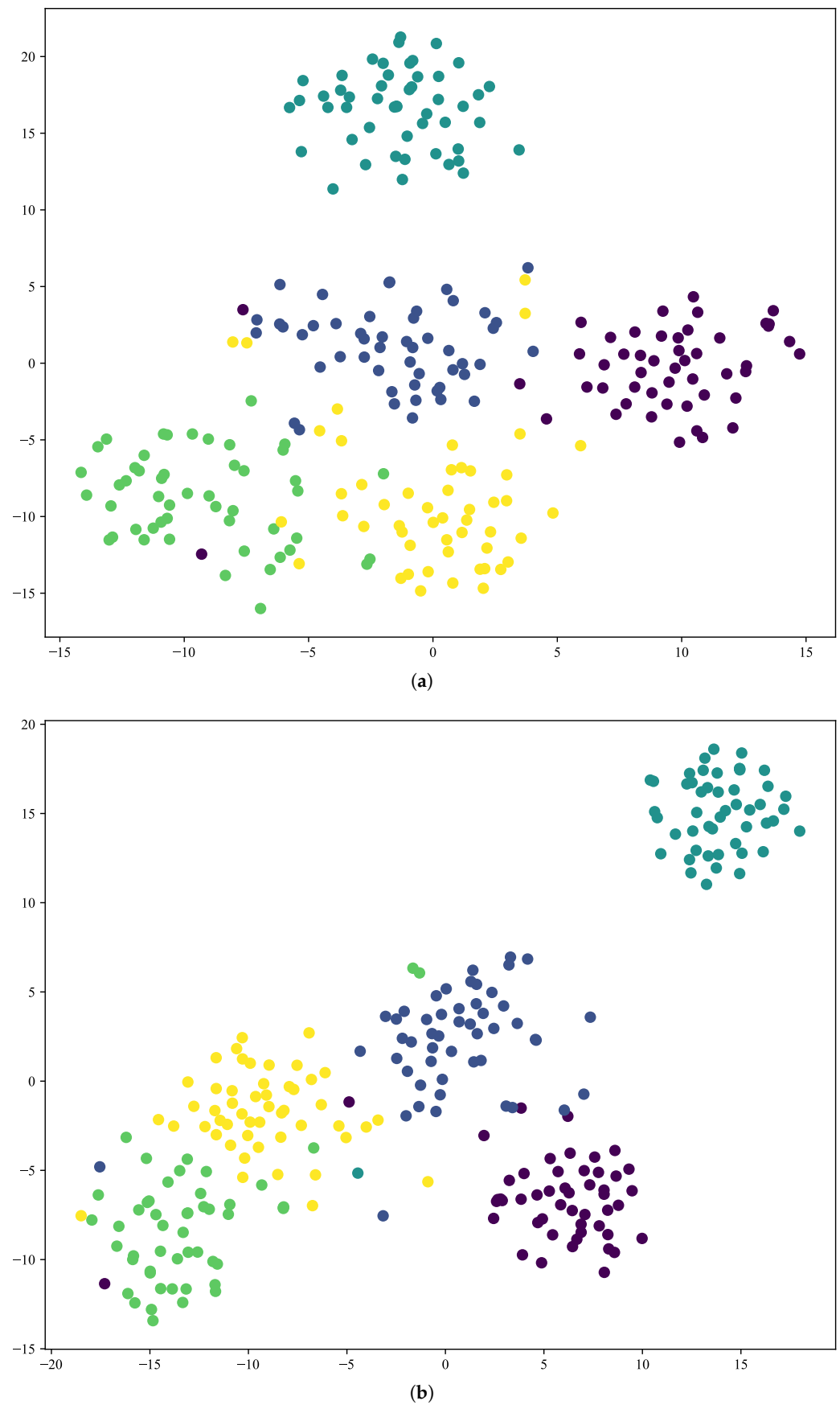


Figure 10. Visualization of feature distributions of global features and global-local feature fusion on CUB-200-2011: (a) using only global features; (b) global-local feature fusion (Ours).

These visualization findings are highly consistent with the quantitative results obtained on Mini-ImageNet and CUB-200-2011, mutually reinforcing the effectiveness of the proposed feature fusion strategy. The results demonstrate that integrating global contextual and local detailed information significantly boosts the model's feature representation capability in few-shot learning scenarios. The dual-path fusion mechanism not only increases the model's sensitivity to class differences, but also improves robustness under complex data distributions by leveraging complementary features across different levels, thus offering a more expressive feature space for few-shot classification tasks.

5. Conclusions

This study addresses the underutilization of global and local features in few-shot image classification by proposing a novel global–local feature fusion algorithm. The method employs a random cropping strategy to obtain multi-scale local views and integrates them through a self-attention mechanism for effective local feature fusion, which is then complementarily combined with global features. This process significantly enhances the model's ability to capture key features and learn more discriminative feature representations. The overall framework of the proposed approach and the design of the self-attention-based local feature fusion module were introduced in detail. This module leverages self-attention to weight and integrate local features, effectively suppressing background noise while strengthening the representation of salient features. Experimental evaluations on the Mini-ImageNet, Tiered-ImageNet, FC-100, and CUB-200-2011 datasets demonstrate that the proposed global–local feature fusion strategy consistently achieves superior performance compared with state-of-the-art few-shot classification methods. A series of ablation studies conducted on the Mini-ImageNet dataset further verified the synergistic effects of the core components. Results show that the feature fusion module, via self-attention, effectively integrates local features, yielding notable classification improvements. The fusion of global features with locally enhanced features provides an additional performance gain of approximately 0.5%, underscoring the complementary strengths of global contextual information and local fine-grained details. The benefits of the feature fusion module are especially pronounced in the 1-shot setting, indicating strong robustness under extreme data scarcity. The proposed feature fusion architecture offers a new and effective approach for optimizing feature representation in few-shot learning, enabling significant improvements in both classification accuracy and generalization capability across multiple benchmark datasets.

Author Contributions: L.Z. is mainly responsible for the supervision and leadership of the planning and implementation of scientific research activities. L.Z. and X.C. are mainly responsible for the research design and code writing and article writing. Y.L. and X.Y. are mainly responsible for literature search and format proofreading. Y.L. and W.C. are mainly responsible for LaTeX typesetting and drawing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011717, Special Projects for Key Fields of the Education Department of Guangdong Province under Grant 2024ZDZX1048.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and analyzed during the current study available from the corresponding author on reasonable request. All data generated or analyzed during this study are included in this article. The code will be open sourced at <https://github.com/DoNotWantToGrowUp> (accessed on 19 September 2025).

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

References

1. Askari, F.; Fateh, A.; Mohammadi, M.R. Enhancing few-shot image classification through learnable multi-scale embedding and attention mechanisms. *Neural Netw.* **2025**, *187*, 107339. [\[PubMed\]](#)
2. Ouahab, A.; Ahmed, O.B. ProtoMed: Prototypical networks with auxiliary regularization for few-shot medical image classification. *Image Vis. Comput.* **2025**, *154*, 105337. [\[CrossRef\]](#)
3. Ren, J.; Li, C.; An, Y.; Zhang, W.; Sun, C. Few-Shot Fine-Grained Image Classification: A Comprehensive Review. *AI* **2024**, *5*, 405–425. [\[CrossRef\]](#)
4. Jaradat, S.; Elhenawy, M.; Nayak, R.; Paz, A.; Ashqar, H.I.; Glaser, S. Multimodal Data Fusion for Tabular and Textual Data: Zero-Shot, Few-Shot, and Fine-Tuning of Generative Pre-Trained Transformer Models. *AI* **2025**, *6*, 72. [\[CrossRef\]](#)
5. Chen, H.; Lindshield, S.; Ndiaye, P.I.; Ndiaye, Y.H.; Pruetz, J.D.; Reibman, A.R. Applying Few-Shot Learning for In-the-Wild Camera-Trap Species Classification. *AI* **2023**, *4*, 574–597. [\[CrossRef\]](#)
6. Dahia, G.; Segundo, M.P. Meta Learning for Few-Shot One-Class Classification. *AI* **2020**, *2*, 195–208. [\[CrossRef\]](#)
7. Ganesan, P.; Jagatheesaperumal, S.K.; Hassan, M.M.; Pupo, F.; Fortino, G. Few-shot image classification using graph neural network with fine-grained feature descriptors. *Neurocomputing* **2024**, *610*, 128448. [\[CrossRef\]](#)
8. Noman, A.; Beiji, Z.; Zhu, C.; Alhabib, M.; Al-Sabri, R. FEGGNN: Feature-Enhanced Gated Graph Neural Network for robust few-shot skin disease classification. *Comput. Biol. Med.* **2025**, *189*, 109902. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Liu, Y.; Zhang, H.; Yang, Y. Few-Shot Image Classification Based on Asymmetric Convolution and Attention Mechanism. In Proceedings of the 2022 4th International Conference on Natural Language Processing, Xi'an, China, 25–27 March 2022; pp. 217–222. [\[CrossRef\]](#)
10. Lin, Y.; Xie, Z.; Chen, T.; Cheng, X.; Wen, H. Image privacy protection scheme based on high-quality reconstruction DCT compression and nonlinear dynamics. *Expert Syst. Appl.* **2024**, *257*, 124891. [\[CrossRef\]](#)
11. Xie, Z.; Xie, W.; Cheng, X.; Yuan, Z.; Cheng, W.; Lin, Y. Image Privacy Protection Communication Scheme by Fibonacci Interleaved Diffusion and Non-Degenerate Discrete Chaos. *Entropy* **2025**, *27*, 790. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Gao, S.; Ding, S.; Iu, H.H.C.; Erkan, U.; Toktas, A.; Simsek, C.; Wu, R.; Xu, X.; Cao, Y.; Mou, J. A three-dimensional memristor-based hyperchaotic map for pseudorandom number generation and multi-image encryption. *Chaos Interdiscip. J. Nonlinear Sci.* **2025**, *35*, 073105. [\[CrossRef\]](#) [\[PubMed\]](#)
13. He, C.; Feng, K.; Zhao, H. Cross-modal Collaboration for Augmented Few-Shot Image Classification. In Proceedings of the 2024 7th International Conference on Information Communication and Signal Processing (ICICSP), Zhoushan, China, 21–23 September 2024; pp. 181–185. [\[CrossRef\]](#)
14. Chen, X.; Lin, P. Few-Shot Image Classification Based on Multimodal Information Fusion. In Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Shenzhen, China, 22–24 November 2024; pp. 657–660. [\[CrossRef\]](#)
15. Wang, Z.; Li, Y.; Zhang, R.; Wang, J.; Cui, H. More diversity, less redundancy: Feature refinement network for few-shot SAR image classification. *Comput. Electr. Eng.* **2025**, *123*, 110043. [\[CrossRef\]](#)
16. Zeng, S.; Xia, Y.; Gu, S.; Liu, F.; Zhou, J. Few-shot classification for soil images: Prototype correction and feature distance enhancement. *Comput. Electron. Agric.* **2025**, *233*, 110162. [\[CrossRef\]](#)
17. Chen, N.; Kuang, X.; Liu, F.; Wang, K.; Zhang, L.; Chen, Q. Few-shot image classification based on gradual machine learning. *Expert Syst. Appl.* **2024**, *255*, 124676. [\[CrossRef\]](#)
18. Zhang, L.; Lin, Y.; Yang, X.; Chen, T.; Cheng, X.; Cheng, W. From Sample Poverty to Rich Feature Learning: A New Metric Learning Method for Few-Shot Classification. *IEEE Access* **2024**, *12*, 124990–125002. [\[CrossRef\]](#)
19. Zhang, S.; Chen, Z.; Zhong, F. Cross-domain multimodal feature enhancement hypergraph neural network for few-shot hyperspectral images classification. *Expert Syst. Appl.* **2025**, *283*, 127742. [\[CrossRef\]](#)
20. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 7260–7268.
21. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
22. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
23. Oreshkin, B.; Rodríguez López, P.; Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018.
24. Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; Wang, X. Meta-baseline: Exploring simple meta-learning for few-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9062–9071.

25. Zhang, C.; Cai, Y.; Lin, G.; Shen, C. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12203–12213.
26. Wu, Z.; Zhao, H. Hierarchical few-shot learning with feature fusion driven by data and knowledge. *Inf. Sci. Int. J.* **2023**, *639*, 119012. [\[CrossRef\]](#)
27. Padmanabhan, D.C.; Gowda, S.; Arani, E.; Zonooz, B. Lsfl: Leveraging shape information in few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4971–4980.
28. Chen, H.; Li, H.; Li, Y.; Chen, C. Sparse spatial transformers for few-shot learning. *Sci. China Inf. Sci.* **2023**, *66*, 210102. [\[CrossRef\]](#)
29. Wang, X.; Wang, X.; Jiang, B.; Luo, B. Few-Shot Learning Meets Transformer: Unified Query-Support Transformers for Few-Shot Classification. *arXiv* **2022**, arXiv:2208.12398. [\[CrossRef\]](#)
30. Lee, K.; Maji, S.; Ravichandran, A.; Soatto, S. Meta-Learning with Differentiable Convex Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
31. Mangla, P.; Kumari, N.; Sinha, A.; Singh, M.; Krishnamurthy, B.; Balasubramanian, V.N. Charting the right manifold: Manifold mixup for few-shot learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2020; pp. 2218–2227.
32. 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.
33. Chen, H.; Li, H.; Li, Y.; Chen, C. Multi-scale adaptive task attention network for few-shot learning. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 4765–4771.
34. Lu, S.; Ye, H.J.; Zhan, D.C. Tailoring embedding function to heterogeneous few-shot tasks by global and local feature adaptors. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 19–21 May 2021; Volume 35, pp. 8776–8783.
35. Li, K.; Zhang, Y.; Li, K.; Fu, Y. Adversarial feature hallucination networks for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13470–13479.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.