Region Partitioning and Prototype-Query Calibration for Few-Shot Fine-Grained Image Classification

1st Zhiqiang Guo
School of Information Engineering
Wuhan University of Technology
Wuhan, China
guozhiqiang@whut.edu.cn

2nd Longgang Xiao
School of Information Engineering
Wuhan University of Technology
Wuhan, China
longgangxiao@whut.edu.cn

3rd Qiwen Jin*
School of Information Engineering
Wuhan University of Technology
Wuhan, China
qiwenjin@whut.edu.cn

***Abstract*—Few-shot fine-grained image classification aims to distinguish sub-classes within the same parent class using only a few labeled support samples. The main challenges arise from severe background interference and significant pose variations. Existing methods often rely on external auxiliary models and fail to mitigate the bias between support and query samples. In this paper, we propose an end-to-end framework that overcomes these challenges by reducing background interference and pose instability. Our approach decouples feature extraction into global and local branches to enhance foreground regions and suppress irrelevant background noise. Query features and prototypes are then calibrated through similarity weighting and cross-attention to ensure robust matching despite distribution shifts. Moreover, we introduce a Region Partition Module (RPM) to better address background noise and a Prototype-Query Calibration Module (PQCM) that stabilizes class prototypes and aligns query features, improving matching accuracy. Extensive experiments on three fine-grained benchmark datasets demonstrate that our model significantly outperforms state-of-the-art methods. Code is available in https://github.com/969553889/TRMModel.git.**

***Index Terms*—Few-shot learning, Fine-grained image classification, Background suppression, Feature calibration.**

I. Introduction

Few-shot learning (FSL) [1]–[3] has become a focal point in computer vision [4], [5] for its ability to enable models to recognize new concepts from minimal labeled data, mimicking human learning. Unlike traditional methods, which rely on large annotated datasets, FSL is well-suited for real-world scenarios with limited data. Fine-grained image classification (FGIC) [6], [7] involves distinguishing between sub-categories within a broader class, which is especially challenging. Annotating fine-grained images requires domain expertise, making large-scale labeled datasets both time-consuming and labor-intensive. Consequently, FGIC is a natural fit for FSL. This paper focuses on the task of few-shot fine-grained image classification (FS-FGIC) [8]–[11], aiming to recognize fine-grained objects with minimal training examples.

However, FS-FGIC still remains challenging because fine-grained categories exhibit high inter-class similarity and discriminative cues are often subtle and localized. Most few-shot methods follow a metric-learning paradigm, where class prototypes are estimated from only a handful of support samples and then used to match query instances. Under such scarce supervision, the learned embeddings are easily influenced by incidental patterns, leading to high-variance and potentially biased prototypes, as well as incomplete coverage of truly discriminative regions.

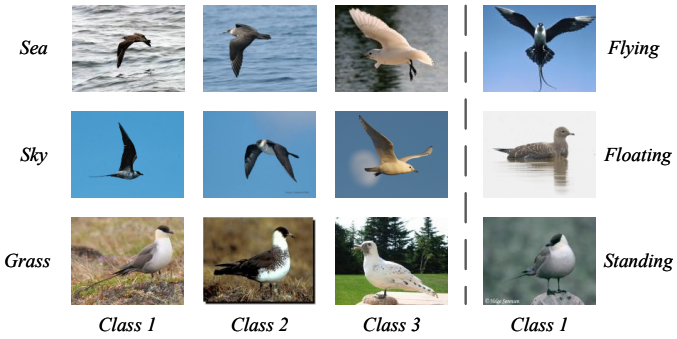

Fig. 1. Illustration of key challenges in few-shot fine-grained image classification: background interference (left), where intra-class samples exhibit diverse backgrounds while inter-class samples may share similar contexts, and pose instability (right), where discriminative regions shift across instances, leading to biased and unstable representations.

Beyond these intrinsic difficulties, FS-FGIC is further challenged by background interference and pose instability, which amplify prototype variance and aggravate support-query distribution shifts. As shown in Fig. 1 (left), images within the same category can have vastly different backgrounds, while different categories may share similar ones. When supervision is limited, models may rely on misleading contextual cues (e.g., textures like water or grass), leading to biased representations. Fig. 1 (right) demonstrates how pose and viewpoint variations can shift discriminative regions, making reliance on a single peak region unstable and prone to missing subtle cues. These issues also lead to a support-query mismatch, where prototypes estimated from limited support instances may be skewed by context or pose, undermining metric matching reliability.

To address these challenges, we propose an end-to-end framework consisting of a Region Partition Module (RPM) and a Prototype-Query Calibration Module (PQCM). The RPM decouples backbone features into global semantic and local metric branches. The global branch generates pixel-wise attention maps to enhance the foreground regions, while the local branch performs peak-background suppression to

*Corresponding author:Qiwen Jin(email:qiwenjin@whut.edu.cn)

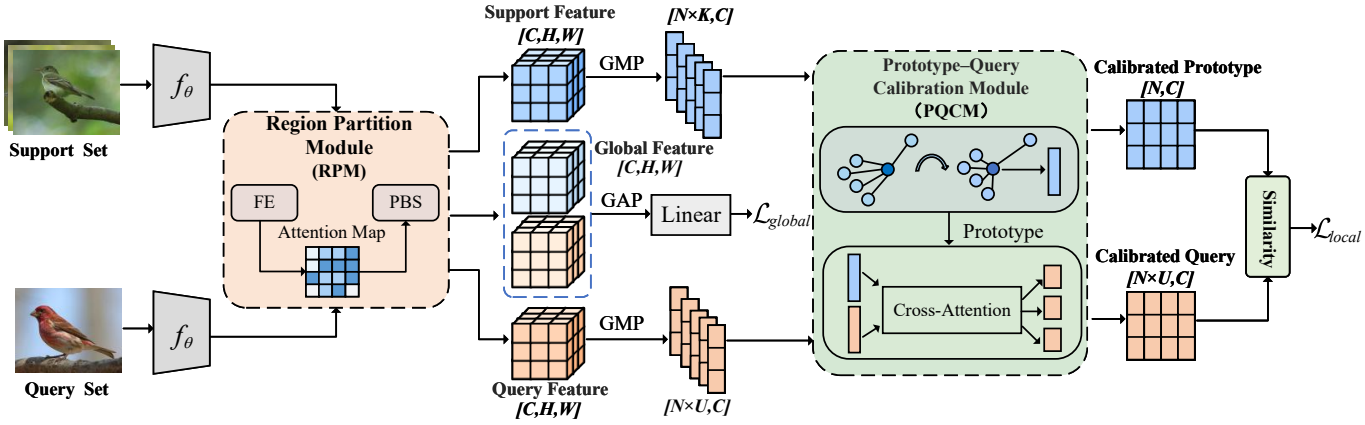

Fig. 2. Overview of the network architecture. The Region Partition Module (RPM) includes Foreground Enhancement (FE), which enhances foreground features, and Peak-Background Suppression (PBS), which reduces background noise. The Prototype-Query Calibration Module (PQCM) calibrates prototypes and query features through similarity weighting and cross-attention, improving robustness against distribution shifts. The model is optimized with global and local loss functions for semantic consistency and metric learning.

reduce irrelevant background noise and highlight more subtle discriminative features. Building on the refined local features, PQCM stabilizes the prototypes by applying similarity-based weighting to reduce the impact of outlier support samples. It then aligns the query features with the calibrated prototypes through cross-attention interaction and gated residual fusion, ensuring reliable matching even in the presence of distribution discrepancies between the support and query sets. Our main contributions can be summarized as follows:

- To reduce background interference and enhance foreground features, we propose the RPM, which separates global and local processing to highlight discriminative regions and suppress noise.
- To address support-query mismatch, we introduce the PQCM, which stabilizes prototypes and ensures reliable matching through similarity weighting and cross-attention.
- We demonstrate the effectiveness of our framework through extensive experiments on three standard FS-FGIC benchmarks, showing consistent performance improvements over existing methods.

II. Related Work

A. Few-Shot Learning

Existing Few-Shot Learning methods are generally categorized into optimization-based meta-learning [12]–[14] and metric-learning [1], [2]. Optimization-based approaches adopt a ”learning-to-learn” paradigm, aiming to acquire a transferable optimization strategy or parameter initialization that facilitates rapid adaptation to novel tasks. Conversely, metric-based methods focus on constructing a discriminative embedding space. These approaches perform classification by measuring the similarity between query and support samples, optimizing the feature space to ensure compact intra-class distributions and separable inter-class margins.

B. Few-Shot Fine-grained Image Classification

Few-Shot Fine-Grained Image Classification tackles the recognition of sub-categories exhibiting subtle inter-class discrepancies and significant intra-class variations under data-scarce regimes. In this challenging setting, irrelevant background clutter often dominates the similarity measurement, obscuring the critical discriminative cues required for fine-grained distinction. Consequently, mitigating background interference has emerged as a pivotal research direction to enhance feature robustness.

To address this, Zha et al. [15] proposed the Background Suppression and Foreground Alignment (BSFA) framework, employing a Background Activation Suppression (BAS) module to generate foreground masks. However, BAS relies on a heuristic mean-value thresholding strategy, which performs rough filtering and may inadvertently suppress discriminative cues. Pursuit of higher precision led Wang et al. [16] and Liu et al. [17] to incorporate external pre-trained saliency detection models (e.g., BASNet [18] or U2-Net [19]) to explicitly identify and mask background regions. While these methods achieve superior performance by leveraging strong saliency priors, their reliance on external auxiliary models inevitably limits the framework's flexibility.

In contrast, we propose the RPM to achieve end-to-end feature partitioning without external detectors. By suppressing both background noise and overly dominant peaks, RPM facilitates the mining of comprehensive semantic details. Furthermore, the PQCM is introduced to align feature distributions, ensuring robust metric learning against intra-class variations.

III. Methods

A. Problem Definition and Overall Framework

The objective of FSL is to acquire transferable knowledge from a base class set $\mathcal{D}_{base}$, enabling the model to perform well on a novel class set $\mathcal{D}_{novel}$ with limited labeled samples, where $\mathcal{D}_{base} \cap \mathcal{D}_{novel} = \emptyset$. Following the standard FSL

protocol, we adopt the episodic training strategy. During the meta-training phase, each episode is formulated as an "N-way K-shot" classification task, where N classes are randomly sampled from $\mathcal{D}_{base}$. For each class, the support set contains K labeled samples, while the query set consists of U samples for evaluation. During the meta-testing phase, the model is evaluated on $\mathcal{D}_{novel}$ using the knowledge distilled during training. This evaluation paradigm measures the model's capability to generalize to unseen categories.

As depicted in Figure 2, the proposed framework comprises a backbone network, a Region Partition Module (RPM), and a Prototype-Query Calibration Module (PQCM). Given the initial features extracted by the backbone, the RPM decouples them into two parallel branches. The global branch employs foreground enhancement for semantic classification and generates a region attention map. Crucially, this map serves as a pixel-wise guide for the metric branch to perform peak-background suppression, enabling the mining of richer features. Subsequently, the refined metric features are processed by the PQCM, which calibrates prototype and query features via Similarity Weighting and Cross-Attention Interaction mechanisms. The entire network is trained end-to-end by jointly optimizing the global classification loss and the local metric loss.

B. Region Partition Module

The Region Partition Module (RPM) is designed to extract comprehensive features through two collaborative mechanisms: Foreground Enhancement (FE) and Peak-Background Suppression (PBS), as illustrated in Figure 3. Both mechanisms share a pre-processing block g_θ that includes two convolutional layers. Let $F \in \mathbb{R}^{C\times H\times W}$ denote the input feature from the backbone, where C is the number of channels, H and W are the height and width of the feature map, respectively. We first refine it to obtain $\dot{F} = g_\theta(F) \in \mathbb{R}^{C\times H\times W}$, which serves as the common input for both branches.

1) Foreground Enhancement: The FE branch primarily aims to construct a robust global representation and generate a precise pixel-wise attention map to guide the subsequent PBS branch. To capture discriminative textures and shapes effectively, we first aggregate channel-wise statistics using Channel-wise Average Pooling (CAP) and Channel-wise K-Max Pooling (CKMP). These statistics are concatenated into a unified descriptor $\hat{F}$. To capture spatial contexts at different scales, this descriptor is processed by two parallel convolutional layers with kernel sizes of 3×3 and 5×5. The outputs of these two branches are then element-wise summed to produce the final attention map $A \in \mathbb{R}^{1\times H\times W}$:

$$\hat{F} = \text{Concat}[\text{CAP}(\dot{F}), \text{CKMP}(\dot{F})], \tag{1}$$

$$A = \sigma\left(\text{Conv}_{3\times3}(\hat{F}) + \text{Conv}_{5\times5}(\hat{F})\right), \tag{2}$$

Here, $\sigma(\cdot)$ denotes the Sigmoid activation function. Utilizing the learned attention map A, we explicitly enhance the foreground regions via a residual connection to obtain the global feature $F_{global} \in \mathbb{R}^{C\times H\times W}$:

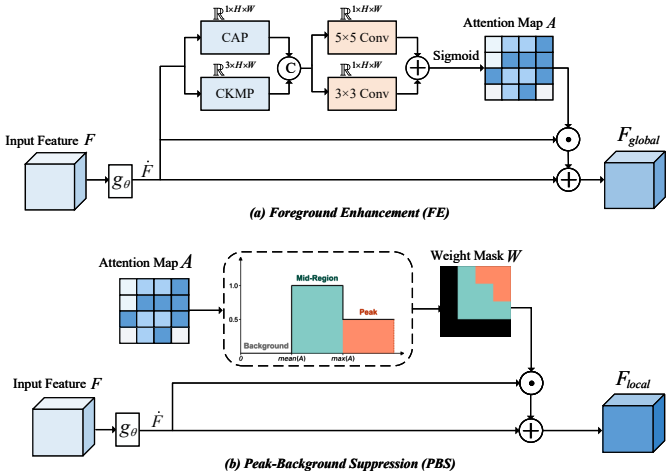

Fig. 3. The structure of Region Partition Module (RPM). RPM consists of a Foreground Enhancement (FE) branch and a Peak-Background Suppression (PBS) branch: FE strengthens foreground-aware responses, while PBS suppresses background noise and over-peaked activations to obtain refined local features.

$$F_{global} = \dot{F} + A \odot \dot{F}. \tag{3}$$

2) Peak-Background Suppression: In fine-grained scenarios, the salient regions of the same species often vary significantly due to diverse backgrounds and poses. Under the few-shot setting with extremely limited data, the model is prone to overfitting to specific prominent patterns or being misled by irrelevant background noise, leading to poor generalization on novel classes. To address this, the model must look beyond the salient peaks and mine more subtle discriminative cues. The PBS branch utilizes the attention map A to partition spatial features into three regions: background, peak, and mid-region. We construct a spatial weight mask W based on the statistical distribution of A:

$$W_{i,j} = \begin{cases} 0, & \text{if } A_{i,j} < mean(A) \\ 0.5, & \text{if } A_{i,j} > \gamma \cdot max(A) \\ 1.0, & \text{otherwise.} \end{cases} \tag{4}$$

where γ is a hyperparameter used to control the degree of suppression.

Specifically, we construct a spatial weight map W to perform relative suppression. We set $W = 0$ on background regions so that these activations receive no residual amplification, while assigning $W = 0.5$ to peak regions to moderate their residual boost. The remaining regions are kept at $W = 1$, allowing informative but less salient cues to be enhanced consistently. In this way, background responses are suppressed in relative importance compared with the amplified foreground regions, and overly dominant peaks are prevented from monopolizing the representation. Finally, the

local feature $F_{local} \in \mathbb{R}^{C\times H\times W}$ is refined through a weighted residual connection:

$$F_{local} = \dot{F} + W \odot \dot{F}. \tag{5}$$

C. Prototype-Query Calibration Module

Although the RPM reduces background interference, the distribution discrepancy between support and query features remains a critical challenge. To address this, we propose the Prototype-Query Calibration Module (PQCM). First, we apply Global Max Pooling (GMP) to the refined local feature maps F_{local} generated by the RPM to obtain feature vectors. These vectors are then organized into a support feature set $F^S = \{f_{i,j}\}_{i=1,j=1}^{N,K}$ and a query feature set F^Q, where N is the number of classes and K is the number of shots. As illustrated in Figure 4, the PQCM processes these features through two sequential steps: Similarity Weighting and Cross-Attention Interaction.

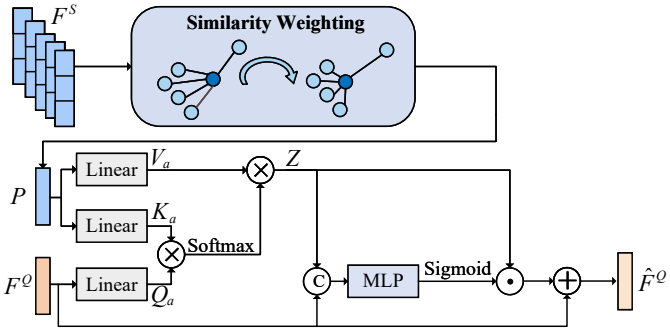

Fig. 4. The structure of Prototype-Query Calibration Module (PQCM). PQCM constructs similarity-weighted prototypes and calibrates query features via cross-attention, yielding robust representations for metric matching.

1) Similarity Weighting: The small size of the support set makes class prototypes highly sensitive to outliers: a single atypical instance can distort the class center and degrade subsequent matching. To dampen this effect, we assign an adaptive weight to each support sample based on its average similarity to the remaining samples of the same class. Specifically, for the i-th class, we assign a confidence score to every support feature $f_{i,j}$. The raw confidence of the j-th sample is defined as the average cosine similarity between this sample and the remaining $K-1$ support samples in the same class:

$$\mu_{i,j} = \frac{1}{K-1}\sum_{\substack{k=1\\k\neq j}}^{K}\cos(f_{i,j}, f_{i,k}), \quad \cos(u,v) = \frac{u^\top v}{\|u\|_2\|v\|_2}, \tag{6}$$

where $\|\cdot\|_2$ represents the L2-norm. These scores are then normalized to form a probability distribution using the Softmax function:

$$\bar{\mu}_{i,j} = \frac{\exp(\mu_{i,j})}{\sum_{k=1}^{K}\exp(\mu_{i,k})}. \tag{7}$$

Using the normalized weights, a robust calibrated prototype P_i for class i is obtained by the weighted summation:

$$P_i = \sum_{j=1}^{K}\bar{\mu}_{i,j}\cdot f_{i,j}. \tag{8}$$

For 1-shot episodes, we set the weight to 1 and the prototype equals the only support feature.

2) Cross-Attention Interaction: After obtaining the calibrated prototypes, denoted as $P \in \mathbb{R}^{N\times C}$, we introduce a Cross-Attention mechanism to align the query features with these prototypes. This step facilitates the transfer of semantic context from the prototypes to the queries, effectively highlighting features relevant to the target classes. We project the query feature set F^Q and the prototype P into the query (Q_a), key (K_a), and value (V_a) subspaces via linear projection. We then compute the attention matrix to capture the semantic correlation between queries and prototypes, which is subsequently used to aggregate the contextual information Z:

$$Z = \text{Softmax}\left(\frac{Q_a K_a^T}{\sqrt{D}}\right)V_a, \tag{9}$$

where D is the number of channels after projection.

To adaptively fuse this context with the original features, we employ a gating mechanism. The original query features F^Q and the aggregated context Z are concatenated and processed by a Multi-Layer Perceptron (MLP) to generate a modulation gate G. The final calibrated query features $\hat{F}^Q$ are produced via a residual connection:

$$G = \sigma\left(\text{MLP}\left(\text{Concat}[F^Q, Z]\right)\right), \tag{10}$$

$$\hat{F}^Q = F^Q + G \odot Z. \tag{11}$$

Following the operations of the RPM and PQCM, we obtain three key components: the global feature F_{global} from the RPM's global branch, which preserves rich semantic information, and the calibrated prototypes P along with the calibrated query features $\hat{F}^Q$ from the PQCM.

D. Objective Function

During the training phase, given an N-way K-shot fine-grained recognition task $\mathcal{T}$, the model processes a batch of samples comprising both the support and query sets. Let $\mathcal{T} = \{x_i\}_{i=1}^{M}$ denote the set of all samples in the current episode, where M is the total number of images (i.e., $M = N(K+U)$). Each sample x_i is associated with two types of labels: a global label $y_i^g \in \{1, 2, ..., |\mathcal{D}_{base}|\}$ corresponding to the actual category in the base set, and a local episodic label $y_i^l \in \{0, 1, ..., N-1\}$ representing the relative index within the current episode. We optimize a joint objective to simultaneously learn generalizable semantic representations and discriminative fine-grained metrics.

To encourage the feature extractor to learn robust and discriminative semantic patterns, we introduce a global classification branch during training. This branch is supervised by a cross-entropy objective over the base-class label space, which enhances inter-class separability and provides stable semantic

TABLE I
PERFORMANCE COMPARISON ON CUB, DOGS, AND CARS DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE UNDER THE SAME BACKBONE.

Method	Backbone	*CUB*		*Dogs*		*Cars*	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet (NeurIPS'17) [1]	Conv-4	51.78 ± 0.93	74.72 ± 0.69	41.60 ± 0.82	56.98 ± 0.75	46.38 ± 0.79	64.58 ± 0.67
Relation (CVPR'18) [2]	Conv-4	60.50 ± 0.51	75.50 ± 0.42	47.35 ± 0.88	66.20 ± 0.74	46.04 ± 0.91	68.52 ± 0.78
FRN (CVPR'21) [20]	Conv-4	69.45 ± 0.49	85.16 ± 0.33	49.37 ± 0.20	67.13 ± 0.17	58.90 ± 0.22	79.65 ± 0.15
OLSA (MM'21) [21]	Conv-4	73.07 ± 0.46	86.24 ± 0.29	55.53 ± 0.45	71.68 ± 0.36	70.13 ± 0.48	84.29 ± 0.31
TOAN (TCSVT'22) [22]	Conv-4	65.34 ± 0.75	80.43 ± 0.60	49.30 ± 0.77	67.16 ± 0.49	65.90 ± 0.72	84.24 ± 0.48
MLSO (TMM'22) [23]	Conv-4	68.21 ± 0.78	82.18 ± 0.47	55.62 ± 0.58	71.98 ± 0.71	67.83 ± 0.63	84.83 ± 0.48
BSFA (TCSVT'23) [15]	Conv-4	59.40 ± 0.97	74.42 ± 0.62	49.13 ± 0.84	63.27 ± 0.73	56.06 ± 0.89	73.28 ± 0.68
BAMM (ESWA'24) [24]	Conv-4	68.55 ± 1.12	84.77 ± 0.79	55.76 ± 1.11	72.59 ± 0.88	63.38 ± 1.09	84.42 ± 0.84
RSaD (TMM'24) [17]	Conv-4	71.15 ± 0.92	84.03 ± 0.62	59.42 ± 0.95	75.30 ± 0.69	65.43 ± 1.29	81.75 ± 0.88
Ours	Conv-4	**74.98 ± 0.49**	**86.56 ± 0.33**	**68.04 ± 0.49**	**82.68 ± 0.30**	**74.15 ± 0.46**	**87.22 ± 0.29**
ProtoNet (NeurIPS'17) [1]	ResNet-12	63.44 ± 0.56	83.17 ± 0.35	41.61 ± 0.50	76.78 ± 0.36	45.01 ± 0.49	87.19 ± 0.31
DN4 (CVPR'19) [25]	ResNet-12	64.95 ± 0.99	83.18 ± 0.62	49.70 ± 0.85	71.59 ± 0.68	75.79 ± 0.84	94.14 ± 0.35
OLSA (MM'21) [21]	ResNet-12	77.77 ± 0.44	89.87 ± 0.24	64.15 ± 0.49	78.28 ± 0.32	77.03 ± 0.46	88.85 ± 0.46
TOAN (TCSVT'22) [22]	ResNet-12	66.10 ± 0.86	82.27 ± 0.60	49.77 ± 0.86	69.29 ± 0.70	75.28 ± 0.72	87.45 ± 0.48
AGPF (PR'22) [26]	ResNet-12	78.73 ± 0.84	89.77 ± 0.47	72.34 ± 0.86	84.02 ± 0.57	85.34 ± 0.74	94.79 ± 0.35
BSFA (TCSVT'23) [15]	ResNet-12	82.22 ± 0.85	90.49 ± 0.47	69.62 ± 0.92	82.50 ± 0.58	**89.42 ± 0.68**	**95.36 ± 0.36**
RSaD (TMM'24) [17]	ResNet-12	82.45 ± 0.79	**92.02 ± 0.44**	73.75 ± 0.93	86.65 ± 0.54	87.27 ± 0.70	95.01 ± 0.49
RaPSPNet (PR'24) [27]	ResNet-12	73.47 ± 0.75	88.57 ± 0.63	61.46 ± 0.82	78.73 ± 0.62	78.84 ± 0.69	93.79 ± 0.49
Ours	ResNet-12	**82.92 ± 0.43**	91.26 ± 0.27	**79.65 ± 0.45**	**89.99 ± 0.25**	88.42 ± 0.36	95.09 ± 0.19

guidance for representation learning. Given the global feature map F^i_{global} produced by the RPM for the i-th training sample, we apply Global Average Pooling (GAP) followed by a linear classifier W_c to predict its global label y^g_i. The global classification loss is defined as:

$$\mathcal{L}_{global} = -\frac{1}{M}\sum_{i=1}^{M}\log\left(\mathrm{Softmax}\big(W_c\,\mathrm{GAP}\big(F^i_{global}\big)\big)_{y^g_i}\right). \tag{12}$$

where $(\cdot)_{y^g_i}$ denotes the predicted probability of the ground-truth class y^g_i. Note that the global branch is only used for training supervision and is discarded during inference, introducing no additional test-time cost.

In parallel, to enable few-shot recognition, we minimize the local metric loss based on the nearest-neighbor classification of query samples. Given the calibrated prototypes P and the calibrated query features $\hat{F}^Q$, the local metric loss is defined as:

$$\mathcal{L}_{local} = -\frac{1}{NU}\sum_{i=1}^{N}\sum_{j=1}^{U}\log\frac{\exp(\tau\cdot\cos(\hat{F}^Q_{i,j},P_i))}{\sum_{k=1}^{N}\exp(\tau\cdot\cos(\hat{F}^Q_{i,j},P_k))}, \tag{13}$$

where τ is a learnable temperature scaling factor.

Finally, the total objective function is a weighted sum of the global semantic loss and the local metric loss:

$$\mathcal{L}_{total} = \lambda\mathcal{L}_{local} + \mathcal{L}_{global}, \tag{14}$$

where λ is a hyperparameter controlling the contribution of the few-shot metric objective. During inference, the global branch is discarded, and recognition relies solely on the metric derived from the calibrated features.

TABLE II
ABLATION STUDY OF DIFFERENT COMPONENTS USING CONV-4 BACKBONE ON CUB AND DOGS DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE.

Baseline	RPM	PQCM	*CUB*		*Dogs*	
			1-shot	5-shot	1-shot	5-shot
✓			69.47	82.76	56.33	71.61
✓	✓		72.09	85.48	65.23	80.84
✓		✓	72.45	85.03	61.67	77.45
✓	✓	✓	**74.98**	**86.56**	**68.04**	**82.68**

TABLE III
ABLATION STUDY OF THE GLOBAL LOSS $\mathcal{L}_{global}$ USING CONV-4 BACKBONE ON CUB AND DOGS DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE.

$\mathcal{L}_{global}$	*CUB*		*Dogs*	
	1-shot	5-shot	1-shot	5-shot
×	72.60 ± 0.53	84.37 ± 0.34	64.63 ± 0.52	79.52 ± 0.33
✓	**74.98 ± 0.49**	**86.56 ± 0.33**	**68.04 ± 0.49**	**82.68 ± 0.30**

IV. EXPERIMENTS

We evaluate our method on three widely recognized benchmarks for few-shot fine-grained image classification: CUB-200-2011 [28], Stanford Dogs [29], and Stanford Cars [30]. For all datasets, we strictly adhere to the data splitting protocol proposed in FRN [20]. CUB-200-2011 (CUB) is a seminal benchmark comprising 11,788 images of 200 bird species. To ensure a strictly fair comparison, we utilize raw images exclusively, avoiding any reliance on human-annotated bounding boxes. Stanford Dogs (Dogs) presents a challenging scenario with 20,580 images across 120 dog breeds. Furthermore, we employ Stanford Cars (Cars), which contains 16,185 images

TABLE IV
ABLATION STUDY OF THE PROPOSED RPM ACROSS DIFFERENT FEW-SHOT LEARNING FRAMEWORKS USING CONV-4 BACKBONE ON CUB AND DOGS DATASETS.

Methods	*CUB*		*Dogs*	
	1-shot	5-shot	1-shot	5-shot
ProtoNet [1]	51.78	74.72	41.60	56.98
ProtoNet+RPM	62.98(+11.20)	81.93(+7.21)	53.25(+11.65)	74.29(+17.31)
Relation [2]	60.50	75.50	47.35	66.20
Relation+RPM	64.92(+4.42)	80.25(+4.75)	54.78(+7.43)	73.86(+7.66)
FRN [20]	69.45	85.16	49.37	67.13
FRN+RPM	73.49(+4.04)	86.18(+1.02)	64.75(+15.38)	80.65(+13.52)

TABLE V
ABLATION STUDY OF REMOVING THE SIMILARITY WEIGHTING IN THE PQCM USING CONV-4 BACKBONE ON CUB AND DOGS DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE.

Similarity Weighting	*CUB*		*Dogs*	
	1-shot	5-shot	1-shot	5-shot
w/o	74.06	85.88	66.46	80.47
w/	**74.98**	**86.56**	**68.04**	**82.68**

representing 196 car models, to extend our validation to rigid objects.

A. Implementation Details

To ensure a fair comparison with state-of-the-art methods, we employ two widely adopted backbones, Conv-4 and ResNet-12, with an input resolution of 84×84. During the training phase, we utilize standard data augmentation techniques including random cropping, horizontal flipping, and color jittering. We optimize the model with SGD, using Nesterov momentum set to 0.9, an initial learning rate of 0.1, and weight decay set to 5e-4. For evaluation, we strictly follow the standard 5-way 1-shot and 5-way 5-shot protocols, selecting the model with the best validation performance for final testing. All reported results are averaged over 2,000 testing episodes with 95% confidence intervals.

B. Comparison with State-of-the-Arts

We compare our proposed method with a number of state-of-the-art methods, including general FSL methods (e.g., ProtoNet [1] and FRN [20]) and specialized FS-FGIC methods (e.g., BSFA [15], RSaD [17], and RaPSPNet [27]). The performance evaluation is conducted on three fine-grained benchmarks, as summarized in Table I.

The upper part of Table I documents the accuracy comparison using the Conv-4 backbone. Our method demonstrates significantly superior performance in the 1-shot condition. Specifically, compared to BSFA, which also employs background suppression and introduces additional global category labeling, our method achieves remarkable improvements of 15.58%, 18.91%, and 18.09% on the CUB, Dogs, and Cars datasets, respectively. The lower section of Table I presents results with the deeper ResNet-12 backbone. Consistent with Conv-4, our method remains highly competitive, particularly in 1-shot scenarios. While we rank slightly below RSaD on CUB (5-shot), this is attributed to its saliency-based masking, which benefits significantly from multiple support samples to refine target regions. Similarly, our slight disadvantage against BSFA on Cars stems from the dataset's rigid, artificial nature, which limits the efficacy of our abstract semantic mining. Nevertheless, across all benchmarks, our method exhibits significantly more stable and consistent state-of-the-art performance.

TABLE VI
THE INFLUENCE OF THE HYPERPARAMETER γ USING CONV-4 BACKBONE ON CUB AND DOGS DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE.

γ	*CUB*		*Dogs*	
	1-shot	5-shot	1-shot	5-shot
0.6	74.22 ± 0.49	85.88 ± 0.33	66.47 ± 0.49	81.27 ± 0.32
0.7	74.40 ± 0.49	86.19 ± 0.32	66.80 ± 0.49	81.77 ± 0.30
0.8	74.85 ± 0.49	86.41 ± 0.32	66.96 ± 0.49	81.72 ± 0.31
0.9	**74.98 ± 0.49**	**86.56 ± 0.33**	**68.04 ± 0.49**	**82.58 ± 0.30**

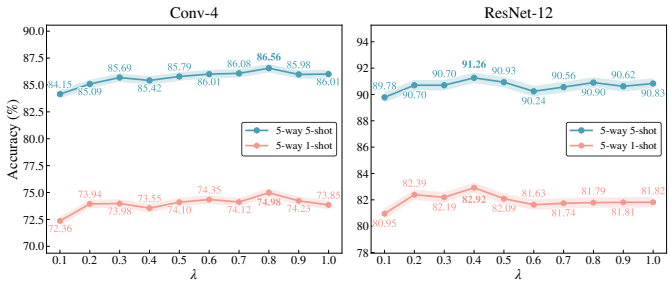

Fig. 5. Sensitivity analysis of the hyperparameter λ on Conv-4 (left) and ResNet-12 (right) backbones. The results indicate that the model performance remains robust across a wide range of λ values.

C. Ablation Experiment

To further validate the effectiveness of the proposed modules and justify the rationale behind their specific design choices, we conduct extensive experiments and analyses.

1) Ablation study of submodules: As shown in Table II, we evaluate the impact of our proposed modules on the CUB and Dogs datasets. Both RPM and PQCM individually improve the baseline, and their combination yields a significant further boost. This validates that RPM enhances semantic mining by suppressing background and peak regions, while PQCM effectively alleviates the support-query distribution bias via dual-end calibration.

2) Ablation study of global loss: To evaluate the contribution of the global loss $\mathcal{L}_{global}$ in our joint optimization, we conduct an ablation study on CUB and Dogs datasets. As shown in Table III, removing $\mathcal{L}_{global}$ leads to consistent performance degradation under both 1-shot and 5-shot settings. This indicates that $\mathcal{L}_{global}$ provides stable semantic supervision during training, encouraging more reliable foreground-related representations and discriminative cues, which further

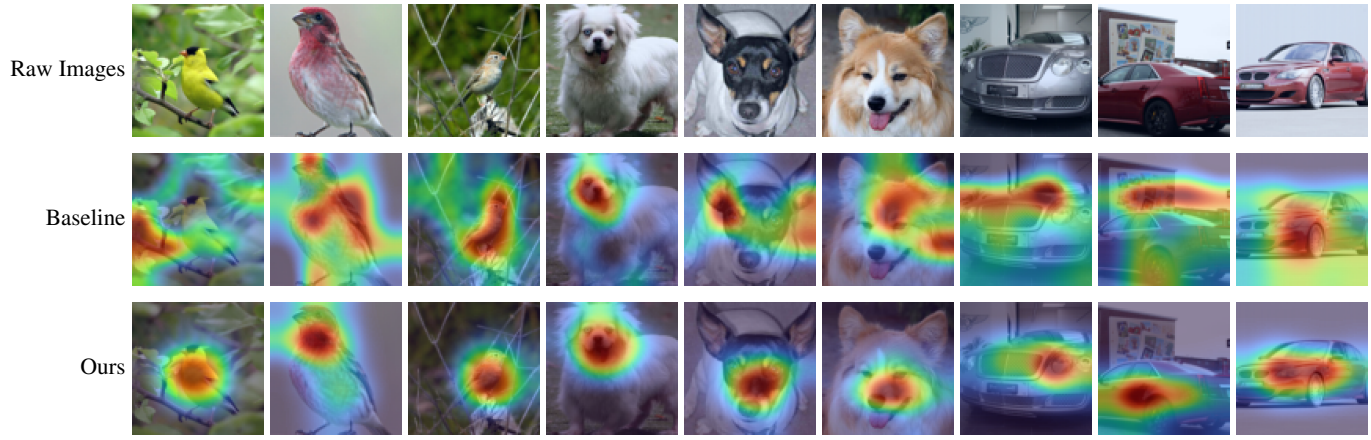

Fig. 6. Visualization of the discriminative regions captured by Baseline and our method (Ours). Our method reduces background distraction and focuses on the most discriminative regions, enabling more reliable fine-grained classification.

benefits the subsequent metric learning and prototype-query matching.

3) Ablation study of RPM: To verify the scalability of RPM, we integrate it into three classic few-shot learning frameworks. RPM consistently improves performance on both CUB and Dogs datasets under 1-shot and 5-shot settings. These gains come from the synergy of its two branches, which emphasize discriminative regions while suppressing background noise, making RPM easy to drop into existing methods.

4) Ablation study of similarity weighting in the PQCM: We conduct ablation experiments with (w/) and without (w/o) similarity weighting. The results, shown in Table V, demonstrate that similarity weighting effectively mitigates the impact of outliers to yield more robust prototypes. Consequently, this facilitates more precise calibration of query features, leading to improved model performance.

5) The influence of λ: To investigate the impact of the hyperparameter λ, we evaluate the model performance on the CUB dataset across various values. As shown in Fig. 5, as λ increases, the classification accuracy exhibits a trend of first rising and then declining. Notably, the optimal λ differs by architecture: the shallower Conv-4 backbone peaks at $\lambda = 0.8$, while the deeper ResNet-12 favors $\lambda = 0.4$. This is because the shallow extractor has a weaker representational capacity and struggles to effectively explore inter-category relationships. Consequently, we set λ to 0.8 and 0.4 for Conv-4 and ResNet-12, respectively.

6) The influence of γ: To investigate the impact of the hyperparameter γ, we evaluate our method with varying values as summarized in Table VI. The results indicate that a small γ leads to excessive suppression of feature maps, causing the loss of valuable discriminative information. In contrast, the model achieves optimal classification accuracy in both 1-shot and 5-shot scenarios when γ is set to 0.9.

D. Visualization

To validate our method, we visualize the discriminative regions in Fig. 6. While the baseline tends to focus on diffuse object outlines or irrelevant background clutter, our method precisely localizes the finest discriminative details, such as bird beaks, dog mouths and car lights. This confirms our model's capability to filter noise and mine robust semantic features.

V. Conclusion

This paper proposes an end-to-end FS-FGIC framework to counter background interference and pose variations that shift discriminative parts, which otherwise amplify prototype variance and intensify support–query distribution shifts. Our method combines a Region Partition Module (RPM) and a Prototype–Query Calibration Module (PQCM): RPM learns pixel-wise attention for foreground enhancement and applies peak–background suppression to filter clutter while encouraging richer local cue mining. PQCM then builds similarity-weighted prototypes and calibrates query features via cross-attention with gated fusion, improving metric matching under distribution mismatch. Experiments on three fine-grained benchmark datasets show the superior classification performance of the proposed method to the state-of-the-art methods.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant no. 62401416 and the Natural Science Foundation of Hubei Province (General Program) under Grant no. JCZRMS202601339.

References

[1] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[2] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[3] M. R. Zafar and N. Khan, "Attentional feature fusion for few-shot learning," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] F. Du, P. Yang, Q. Jia, F. Nan, X. Chen, and Y. Yang, "Global and local mixture consistency cumulative learning for long-tailed visual recognitions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 814–15 823.

[6] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1309–1322, 2017.

[7] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8927–8948, 2021.

[8] J. Wu, D. Chang, A. Sain, X. Li, Z. Ma, J. Cao, J. Guo, and Y.-Z. Song, "Bi-directional feature reconstruction network for fine-grained few-shot image classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 3, 2023, pp. 2821–2829.

[9] Z.-X. Ma, Z.-D. Chen, L.-J. Zhao, Z.-C. Zhang, X. Luo, and X.-S. Xu, "Cross-layer and cross-sample feature optimization network for few-shot fine-grained image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4136–4144.

[10] J. Sun, Y. Sun, X. Shen, and Q. Sun, "Query-guided prototype optimization for few-shot classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[11] J. Jiang, J. Li, and J. Lu, "View-robust backbone and discriminative reconstruction for few-shot fine-grained image classification," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[13] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 657–10 665.

[14] Y. Zhu, C. Liu, S. Jiang *et al.*, "Multi-attention meta learning for few-shot fine-grained image recognition." in *IJCAI*. Beijing, 2020, pp. 1090–1096.

[15] Z. Zha, H. Tang, Y. Sun, and J. Tang, "Boosting few-shot fine-grained recognition with background suppression and foreground alignment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3947–3961, 2023.

[16] C. Wang, S. Song, Q. Yang, X. Li, and G. Huang, "Fine-grained few shot learning with foreground object transformation," *Neurocomputing*, vol. 466, pp. 16–26, 2021.

[17] H. Liu, C. P. Chen, X. Gong, and T. Zhang, "Robust saliency-aware distillation for few-shot fine-grained visual recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 7529–7542, 2024.

[18] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489.

[19] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.

[20] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8012–8021.

[21] Y. Wu, B. Zhang, G. Yu, W. Zhang, B. Wang, T. Chen, and J. Fan, "Object-aware long-short-range spatial alignment for few-shot fine-grained image classification," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 107–115.

[22] H. Huang, J. Zhang, L. Yu, J. Zhang, Q. Wu, and C. Xu, "Toan: Target-oriented alignment network for fine-grained image categorization with few labeled samples," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 853–866, 2021.

[23] H. Zhang, H. Li, and P. Koniusz, "Multi-level second-order few-shot learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 2111–2126, 2022.

[24] Y. Wang, Y. Ji, W. Wang, and B. Wang, "Bi-channel attention meta learning for few-shot fine-grained image recognition," *Expert Systems with Applications*, vol. 242, p. 122741, 2024.

[25] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7260–7268.

[26] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognition*, vol. 130, p. 108792, 2022.

[27] W. Zhang, Y. Zhao, Y. Gao, and C. Sun, "Re-abstraction and perturbing support pair network for few-shot fine-grained image classification," *Pattern Recognition*, vol. 148, p. 110158, 2024.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[29] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, no. 1, 2011.

[30] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.