ADAPTIVE GLOBAL-LOCAL BALANCE NETWORK FOR FEW-SHOT FINE-GRAINED IMAGE CLASSIFICATION

Yongqi Liu, Yao Chen*, Zhi-Jie Wang†*

College of Computer Science, Chongqing University

ABSTRACT

Few-shot fine-grained image classification (FSFGIC) confronts the dual challenges of limited samples and subtle inter-class differences. Current metric-learning approaches often inadequately balance the need for stable global representations and discriminative local features, leading to unsatisfactory performance. To address this, we propose an Adaptive Global-Local Balance Network (AGLB-Net), a framework that dynamically balances global and local feature representation to achieve robust performance in FSFGIC. Our method introduces two key modules: a Hierarchical Discriminative Feature Refinement (HDFR) module, which progressively integrates context from global semantics to fine-grained details, and an Adaptive Regional Re-Attention Module (ARRM), which automatically localizes and amplifies discriminative regions without additional supervision. Extensive experiments on widely-used benchmarks demonstrate that AGLB-Net consistently achieves state-of-the-art performance across various few-shot settings, validating its competitiveness in FSFGIC.

***Index Terms*—** Few-shot learning, Fine-grained image classification.

1. INTRODUCTION

Image classification has received much attention due to its wide applications [1]. In this community, two particularly active topics are few-shot image classification (FSIC) [2] and fine-grained image classification (FGIC) [3].

FSIC aims to recognize new image classes with only a few labeled samples. The extreme data scarcity hampers model training, resulting in poor generalization to novel classes [4]. To address this, metric-based meta-learning has become the dominant paradigm in the field. Pioneering works like ProtoNet [5] introduced a metric-based approach that classifies query samples by measuring their distances to prototype representations of each class in a learned embedding space. Following this spirit, numerous methods have been developed to handle more sophisticated distance metrics [6, 7], and reconstruction-based approaches like FRN [8] replace explicit metric comparison with feature reconstruction, classifying queries based on their ability to be reconstructed from support features..

On the other hand, FGIC requires distinguishing highly similar subclasses, demanding a strong capability to capture subtle, discriminative features [9]. Early methods often relied on external information like bounding boxes or part annotations [10, 11]. More recent research has shifted toward annotation-free paradigms that focus on designing modules to automatically localize and amplify critical regions without manual guidance [12, 13].

Particularly, recent efforts have turned to few-shot fine-grained image classification (FSFGIC), a domain that bears the challenges from both FSIC and FGIC [14, 15]. To address this "combined" problem, FSFGIC requires models to capture fine-grained details under extreme low-data conditions. In response, the field has largely extended the metric-learning paradigm from few-shot learning, designing models that discriminate based on query-support similarity [16, 17].

However, this direct adoption of the metric-learning paradigm fails to resolve a fundamental contradiction inherent to FSFGIC. Generally speaking, effective fine-grained discrimination necessitates a focus on localized, subtle cues to capture minute inter-class variances; yet, under the few-shot setting, these subtle cues are scarce and noisy, making them prone to overfitting and thus unreliable, whereas global representations—while lacking discriminative power—provide a more stable option. Consequently, merely refining the similarity measurement or feature reconstruction—without addressing what information to prioritize—cannot overcome this core contradiction.

To reconcile this conflict, we propose the **Adaptive Global-Local Balance Network (AGLB-Net)** that adaptively balances global and local information, maximizing the benefits of each while mitigating their drawbacks in FSFGIC, respectively. Specifically, to simultaneously preserve global semantics and exploit fine-grained discriminative cues, we introduce the **Hierarchical Discriminative Feature Refinement (HDFR)** module. HDFR employs a multi-scale bi-directional cross attention mechanism that progressively captures information from global, local, to detail-level granularity and fuses these heterogeneous representations into a unified embedding that is both globally stable and locally discriminative. Furthermore, we present the **Adaptive Regional Re-Attention Module (ARRM)**, which further zooms in on fine-grained details. ARRM dynamically crops the most informative regions guided by learned attention map, introducing an aggressive yet controllable focus on subtle inter-class differences. The outputs of HDFR and ARRM are independently forwarded to a metric-learning classifier to compute query–support distances. A learnable dynamic weight then adaptively balances the contributions of the two modules, explicitly balancing the influence of local feature and global feature during classification.

To summarize, our main contributions can be threefold:

- We propose the Adaptive Global-Local Balance Network (AGLB-Net), a novel framework designed to dynamically balance global semantics and fine-grained local features in few-shot fine-grained image classification.
- We introduce the Hierarchical Discriminative Feature Refinement (HDFR) for multi-scale feature fusion and the Adaptive Regional Re-Attention Module (ARRM) for automatic discriminative region highlighting, enhancing both representation stability and subtle difference capture.

* denotes equal contribution. † denotes corresponding author (Zhi-Jie Wang, cszjwang@cqu.edu.cn).

- Extensive experiments on widely-used benchmarks demonstrate that our network, combined with different metric-learning methods, achieves state-of-the-art performance.

2. METHOD

2.1. Problem Formulation

Following previous works [18], we denote a dataset as $D = \{(x_i, y_i) \mid y_i \in Y\}$, where x_i and y_i are the feature vector and class label of the i-th image, respectively. The label space Y is split into disjoint subsets Y_{train}, Y_{val}, and Y_{test}, corresponding to the training set D_{train}, validation set D_{val}, and test set D_{test}, with $Y_{train} \cup Y_{val} \cup Y_{test} = Y$. A FSIC task is defined as an n-way k-shot episode. Specifically, n classes are randomly sampled from one of the label subsets, and each class provides k labeled samples (support set S) and u unlabeled samples (query set Q), i.e., $|S| = n \times k$, $|Q| = n \times u$, with $S \cap Q = \emptyset$. The goal is to predict the labels of the query set based on the support set.

2.2. Overview

As illustrated in Figure 1, we present the overall framework of the AGLB-Net. Input images are processed by a backbone network. To balance complexity and representational power, we extract hierarchical features from the last two layers of the backbone: the final layer feature F_h and the penultimate layer feature F_m. The core Hierarchical Discriminative Feature Refinement (HDFR) module, taking F_h as its input, enables bidirectional interaction between support and query features, generating multi-scale enhanced representations. Concurrently, an Adaptive Regional Re-Attention Module (ARRM) processes feature F_m to identify and refine crucial local regions guided by attention map. Finally, representations from both the global-context-aware HDFR and the locally-focused ARRM are integrated for metric-based classification.

2.3. Hierarchical Discriminative Feature Refinement

High-level semantics from a single scale are insufficient to discriminate fine-grained categories with subtle inter-class variations. Furthermore, due to the restricted training set in FSIC, the extracted representations may overfit irrelevant details, leading to poor generalization across fine-grained categories. To mitigate this, we design a module, called the Hierarchical Discriminative Feature Refinement (HDFR) module, to extract multi-scale representations from the backbone and reinforce their discriminative power.

The HDFR module performs a direct sequential refinement of support (s) and query (q) features across three progressively focused stages: **global** $\rightarrow$ **local** $\rightarrow$ **detail**. Let $F_{h,s}^{(0)}, F_{h,q}^{(0)} \in \mathbb{R}^{H \times W \times C}$ denote the initial support and query features, respectively. Each stage consists of (i) bi-directional cross-attention, and (ii) gate fusion. Finally, the outputs of all three stages are concatenated with the original input and fused by a 1×1 convolution (channel reduction by 4) and normalization to produce the final refined features. We now proceed to the details of each stage.

Bi-Directional Cross-Attention: For each stage $p \in \{1, 2, 3\}$ (1: global, 2: local, 3: detail), we apply a stage-specific convolutional extractor $\Phi_p(\cdot)$, the output $\hat{F}_{h,s}^{(p)}$ and $\hat{F}_{h,q}^{(p)}$ are computed as:

$$\hat{F}_{h,s}^{(p)} = \Phi_p\left(F_{h,s}^{(p-1)}\right), \quad \hat{F}_{h,q}^{(p)} = \Phi_p\left(F_{h,q}^{(p-1)}\right) \tag{1}$$

In practice, each stage employs a cascade of convolutions with progressively smaller kernels: Φ_1 uses a $5 \times 5 \rightarrow 3 \times 3$ sequence, Φ_2 uses two 3×3 convolutions, and Φ_3 uses a $3 \times 3 \rightarrow 1 \times 1$ sequence, forming stage-specific feature extractors.

Let $L = H \cdot W$. We first reshape the features $\hat{F}_{h,s}^{(p)}$ and $\hat{F}_{h,q}^{(p)}$ into sequences of size $L \times C$, and then map them into queries ($Q^{(p)}$), keys ($K^{(p)}$), and values ($V^{(p)}$) via a 1×1 convolution layer. The (query-from-support) cross-attention update $\bar{F}_{h,q}^{(p)}$ is then computed as:

$$\bar{F}_{h,q}^{(p)} = \mathrm{Softmax}\left(\frac{Q_q^{(p)}(K_s^{(p)})^T}{\sqrt{d_k}}\right) V_s^{(p)} \tag{2}$$

where the support-from-query update $\bar{F}_{h,s}^{(p)}$ is computed symmetrically by swapping the roles of s and q and d_k is the number of keys' channels.

Gate Fusion: Each stage fuses its own input $F_h^{(p-1)}$(output from the previous stage) and attention-enhanced feature $\bar{F}_h^{(p)}$ via a learned gate $G^{(p)}$:

$$G^{(p)} = \sigma\Big(\Psi\big([F_h^{(p-1)}, \bar{F}_h^{(p)}]\big)\Big) \tag{3}$$

where $\Psi(\cdot)$ denotes two successive 1×1 convolutional layers and $\sigma(\cdot)$ is the sigmoid activation. The fused feature $F_h^{(p)}$ is then obtained by weighting $F_h^{(p-1)}$ and $\bar{F}_h^{(p)}$ with $G^{(p)}$ and $(1 - G^{(p)})$ respectively. This operation is applied to both support and query branches.

Final Integration: The hierarchical features from the three stages are concatenated together with the original input feature $F_h^{(0)}$, forming a combined representation F_h^{cat} with channel depth $4C$:

$$F_h^{cat} = \mathcal{C}\left[F_h^{(0)},\, F_h^{(1)},\, F_h^{(2)},\, F_h^{(3)}\right] \tag{4}$$

This feature set is then channel-reduced via a 1×1 convolution, normalized with Layer Norm($LN(\cdot)$), and finally integrated through a residual connection with a learnable scaling factor γ:

$$F^{\mathrm{H}} = F_h^{(0)} + \gamma \cdot \mathrm{LN}\left(\mathrm{Conv}_{1\times1}\left(F_h^{cat}\right)\right) \tag{5}$$

The same procedure is applied consistently to both the support and query branches.

2.4. Adaptive Regional Re-Attention Module

To further exploit fine-grained cues, we propose the Adaptive Regional Re-Attention Module (ARRM). Unlike HDFR, which hierarchically refines features across scales, ARRM adaptively identifies and emphasizes the most informative regions within an image. It operates on the penultimate feature map under the guidance of HDFR and proceeds through four steps: (i) bi-directional cross-attention, (ii) spatial attention fusion with HDFR guidance, (iii) sliding-window region selection, and (iv) re-attention on the cropped region. By concentrating on adaptively selected regions, ARRM amplifies subtle inter-class differences that are easily overlooked in few-shot settings. This targeted attention complements the globally stabilized representations produced by HDFR, yielding a refined, region-aware embedding for fine-grained classification. The following subsection details the four steps of ARRM.

Bi-Directional Cross-Attention: The penultimate layer feature $F_m \in \mathbb{R}^{H' \times W' \times C'}$ is enhanced using the same bi-directional cross-attention mechanism introduced in HDFR (Eq. (2)) and the convolutional settings are the same as the detail stage, yielding the refined feature $\bar{F}_m$.

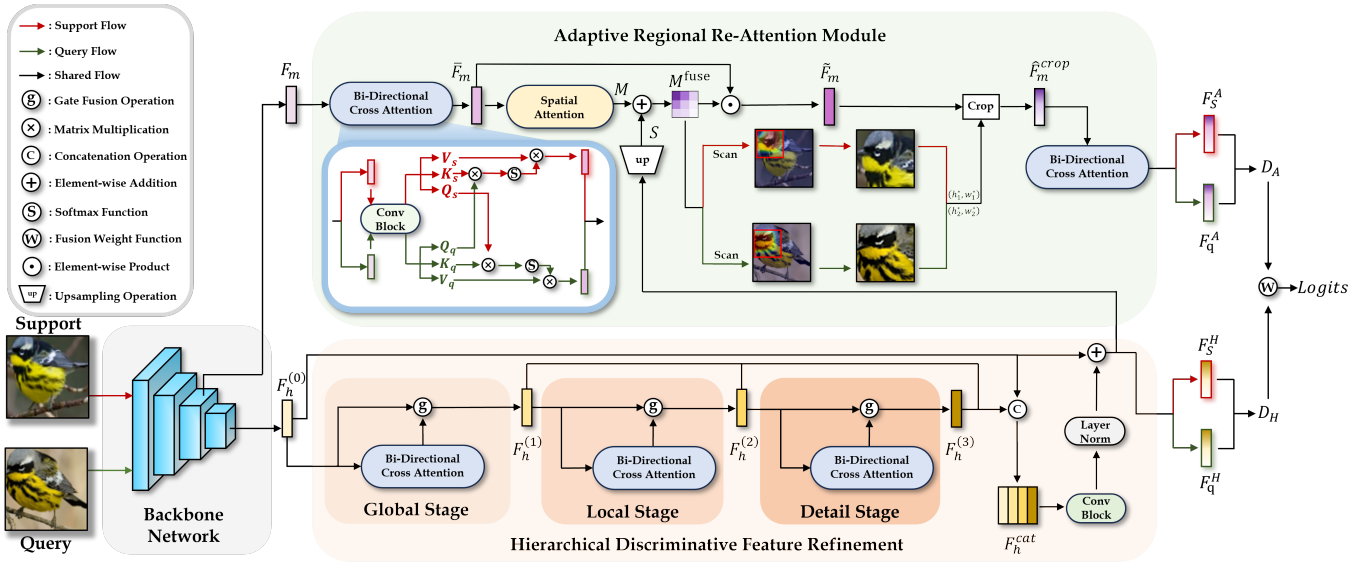

Fig. 1: Overview of the Adaptive Global-Local Balance Network (AGLB-Net),consisting of the Backbone Network, the Hierarchical Discriminative Feature Refinement (HDFR), the Adaptive Regional Re-Attention Module (ARRM) and the final metric classification.

Spatial Attention Fusion with HDFR Guidance: An internal spatial attention map $M = \mathcal{A}(\bar{F}_m) \in \mathbb{R}^{H' \times W'}$ is first computed using a channel-compression convolutional block $\mathcal{A}(\cdot)$. Simultaneously, the HDFR output F^{H} is first reduced in channel dimension, processed by a 1×1 convolutional layer with sigmoid activation, and upsampled to produce the semantic guidance map $S \in \mathbb{R}^{H' \times W'}$. The two attention signals are fused through a learnable parameter β:

$$M^{\mathrm{fuse}} = \sigma(\beta \cdot M + (1 - \beta) \cdot S) \tag{6}$$

where σ denotes the sigmoid function. This fused map M^{fuse} guides next region selection. We obtain $\widetilde{F}_m$ by applying M^{fuse} to $\bar{F}_m$ via element-wise product.

Sliding-Window Region Selection: Given a crop ratio r (set to 0.5 in all experiments), the target crop size is defined as $h_c = \lfloor rH' \rfloor$, $w_c = \lfloor rW' \rfloor$. A sliding window traverses M^{fuse} of size $H' \times W'$, where each window position is defined by its top-left corner coordinates (h_1, w_1) with $h_1 \in [0, H' - h_c]$ and $w_1 \in [0, W' - w_c]$, and bottom-right corner coordinates $(h_2, w_2) = (h_1 + h_c, w_1 + w_c)$. The region with the highest cumulative attention score is selected by evaluating all possible window positions:

$$(h_1^\star, w_1^\star) = \arg\max_{h_1, w_1} \sum_{i=h_1}^{h_2-1} \sum_{j=w_1}^{w_2-1} M^{\mathrm{fuse}}(i, j) \tag{7}$$

The corresponding bottom-right corner is then given by $h_2^\star = h_1^\star + h_c$ and $w_2^\star = w_1^\star + w_c$. Finally, the region feature $\hat{F}_m^{\mathrm{crop}}$ is cropped from $\widetilde{F}_m$ using the coordinates $(h_1^\star, w_1^\star)$ and $(h_2^\star, w_2^\star)$.

Re-Attention on the Cropped Region: The cropped feature $\hat{F}_m^{\mathrm{crop}}$ is then processed by another bi-directional cross-attention block, identical to the first in structure, to produce the final output F^{A} for metric-based classification.

2.5. Metric Classification

F^{H} and F^{A}, the outputs of the HDFR and ARRM, are fed into two separate but structurally identical metric classifiers (e.g., ProtoNet or FRN), producing distance matrices $\mathbf{D}_H, \mathbf{D}_A \in \mathbb{R}^{B \times N}$, where B denotes the number of samples and N is the number of classes. Write the i-th row (the distances for sample i) as $\mathbf{D}_b^{(i)} \in \mathbb{R}^N$ for branch $b \in \{H, A\}$. The probability vector $\mathbf{p}_b^{(i)} = \mathcal{S}(-\mathbf{D}_b^{(i)})$ is obtained by applying a row-wise softmax $\mathcal{S}(\cdot)$ to the negative distances. The entropy for each branch is then computed as:

$$H_b^{(i)} = -\sum_{j=1}^{N} p_{b,j}^{(i)} \log p_{b,j}^{(i)} \tag{8}$$

This entropy $H_b^{(i)}$ measures the uncertainty of branch b on sample i (smaller $H_b^{(i)}$ indicates higher discriminative power). The fusion weight $w_b^{(i)}$ for each branch is computed as the reciprocal of its entropy, normalized by the sum of reciprocals of both branches' entropies (i.e., $w_b^{(i)} \propto 1/H_b^{(i)}$). The final fused distance is obtained by the weight sum:

$$\mathbf{D}_{\mathrm{final}}^{(i)} = \sum_{b \in \{H, A\}} w_b^{(i)} \, \mathbf{D}_b^{(i)} \tag{9}$$

All operations above are row-wise (indexed by i), the superscript (i) denotes the i-th sample's distance vector across classes. The final classification is obtained by applying the log-softmax function to the fused distance vector:

$$\hat{\mathbf{Logits}}^{(i)} = \log \mathrm{softmax}(-\mathbf{D}_{\mathrm{final}}^{(i)}) \tag{10}$$

3. EXPERIMENTS

3.1. Dataset and Implementation Details

We conduct our experiments on three popular FSFGIC benchmarks (CUB200-2011 [19], Stanford Dogs [20], and Stanford Cars [21]). Following the paradigm [16,22] in this field, we divided datasets into D_{train}, D_{val} and D_{test} and all images are resized to 84×84. In our experiments, two widely used backbone architectures (ResNet-12 and Conv-4) are adopted. All experiments are implemented on

an NVIDIA RTX 4090 GPU using PyTorch. We trained models for 900 epochs using SGD with Nesterov momentum of 0.9. The initial learning rate was set to 0.1 and the weight decay to 5×10^{-4}. The learning rate was reduced by a factor of 10 every 300 epochs. Standard data augmentation techniques, including random cropping, random horizontal flipping, and color jittering, were employed to enhance training stability. We conducted evaluation under both 5-way 1-shot and 5-way 5-shot settings, with 15 query images per class. Accuracy scores and 95% confidence intervals are obtained over 2000 trials, as in [23, 24].

3.2. Comparing with State-Of-The-Art

Table 1: 5-way few-shot classification performance of SOTA methods on three benchmark datasets under ResNet-12.

Method	CUB		Dogs		Cars	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [5]	81.14±0.44	91.86±0.24	73.81±0.48	87.39±0.27	85.04±0.42	93.94±0.21
FRN [8]	83.68±0.42	92.89±0.21	76.76±0.47	88.74±0.27	86.03±0.40	95.32±0.17
OLSA [25]	77.77±0.44	89.87±0.24	64.15±0.49	78.28±0.32	77.03±0.46	88.85±0.46
TDM+FRN [26]	83.26±0.20	92.80±0.11	75.98±0.22	88.70±0.13	86.91±0.17	96.11±0.07
C2-Net [23]	84.17±0.42	92.55±0.23	75.50±0.49	87.65±0.28	**88.96±0.37**	95.16±0.20
CSCAM+ProtoNet [27]	81.69±0.20	91.10±0.11	-	-	85.73±0.18	93.26±0.10
CSCAM+FRN [27]	84.00±0.18	93.52±0.10	-	-	86.24±0.18	95.55±0.08
ATR-Net [28]	81.37±0.42	91.26±0.24	75.68±0.46	87.77±0.26	85.60±0.38	95.24±0.16
BSFA [24]	82.27±0.46	90.76±0.26	69.58±0.50	82.59±0.33	88.93±0.38	95.20±0.20
HFCR-Net [29]	83.49±0.19	93.40±0.11	77.01±0.22	88.85±0.13	87.40±0.17	95.88±0.09
Ours+ProtoNet	83.00±0.42	92.60±0.23	75.64±0.47	87.97±0.27	86.31±0.39	94.76±0.20
Ours+FRN	**85.18±0.39**	**94.30±0.20**	**77.21±0.46**	**89.28±0.26**	87.45±0.39	**96.62±0.14**

As shown in Table 1, we conducted 5-way few-shot experiments on the three aforementioned benchmark datasets and compared our method against current state-of-the-art models using ResNet-12 as the backbone. This comparison included two metric learning baselines (ProtoNet and FRN) that can be seamlessly integrated with our method. The experimental results demonstrate that our model exhibits strong performance under both the 5-way 1-shot and 5-way 5-shot settings. When combined with ProtoNet, our model yielded a significant performance improvement over the standalone ProtoNet. Furthermore, when integrated with FRN, our model achieved the highest performance on 5 out of the 6 evaluated metrics. These findings substantiate the powerful performance and stability of our proposed method, confirming its ability to enhance the capabilities of metric learning approaches.

3.3. Ablation Study

Table 2: Ablation study of ARRM and HDFR on CUB-200-2011 dataset, using Protonet and FRN under Conv-4.

		Protonet		FRN	
ARRM	HDFR	1-shot	5-shot	1-shot	5-shot
✗	✗	62.98	83.59	74.27	86.68
✗	✓	75.09	87.95	78.43	90.30
✓	✗	69.24	85.44	77.05	90.14
✓	✓	**76.14**	**88.77**	**78.99**	**90.92**

Table 2 presents the ablation results. And the results demonstrate that combining ARRM and HDFR modules achieves the best performance in both 1-shot and 5-shot tasks, significantly outperforming the baselines. Specifically, in the ProtoNet framework, the full model improves 13.16% over the baseline in 1-shot task. The HDFR module alone contributes significantly to performance, highlighting its role in fine-grained distinctions through multi-scale fusion. When integrated with the more advanced FRN classifier, the ARRM module exhibits better compatibility, especially in the 5-shot scenario, where the combined system reaches 90.92% accuracy. These results confirm that HDFR enhances representational capacity, while ARRM improves discriminative region localization, both improving few-shot fine-grained classification performance.

3.4. Visualization

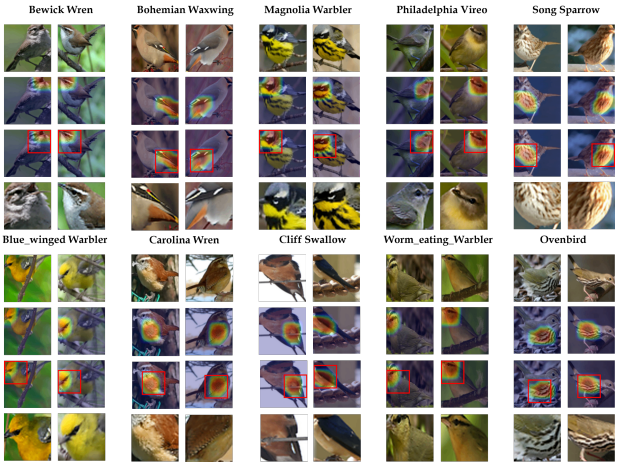

Fig. 2: Visualization of attention map M^{fuse} on the CUB-200-211 dataset.

To illustrate the effectiveness of our cropping mechanism, we visualize the attention maps and corresponding crop regions using the ResNet-12 backbone, the results are shown in Figure 2. There are ten categories of birds, with two samples per category. Each column displays four visualization rows per sample: original image, attention heatmap overlay, cropped region annotation (red box), and zoomed-in cropped region. Critically, our attention mechanism consistently localizes class-discriminative regions in intra-class images, even in the presence of pose and viewpoint variations, allowing the cropped regions to accurately cover these key features. This consistency indicates that our cropping mechanism can adaptively capture category-specific local features, providing a reliable foundation for fine-grained classification.

4. CONCLUSION

This paper presents AGLB-Net, a novel framework that addresses the global-local feature dilemma in FSFGIC through adaptive multi-scale feature balancing. Our method introduces two key modules—HDFR for hierarchical feature refinement and ARRM for discriminative region emphasis—that work together to achieve robust feature representation. Extensive experiments demonstrate state-of-the-art performance across various datasets, validating the effectiveness of our approach in handling both data scarcity and fine-grained distinctions.

5. REFERENCES

[1] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey toward the defender's perspective," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 8:1–8:38, 2023.

[2] J. Y. Lim, K. M. Lim, C. P. Lee, and Y. X. Tan, "A review of few-shot image classification: Approaches, datasets and research trends," *Neurocomputing*, vol. 649, pp. 130774, 2025.

[3] Y. Xie, Q. Gong, X. Luan, J. Yan, and J. Zhang, "A survey of fine-grained visual categorization based on deep learning," *J. Syst. Eng. Electron.*, vol. 35, pp. 1337–1356, 2024.

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, pp. 1126–1135.

[5] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.

[6] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *ICLR*, 2018.

[7] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*, 2020, pp. 12200–12210.

[8] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *CVPR*, 2021, pp. 8008–8017.

[9] X.-S. Wei, Y.-Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 8927–8948, 2022.

[10] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *ECCV*, 2014, pp. 834–849.

[11] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked cnn for fine-grained visual categorization," in *CVPR*, 2016, pp. 1173–1182.

[12] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017, pp. 4476–4484.

[13] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.

[14] J. M. Lim, K. M. Lim, C. P. Lee, and J. Y. Lim, "A review of few-shot fine-grained image classification," *Expert Syst. Appl.*, vol. 275, pp. 127054, 2025.

[15] Y. Liu, Y. Bai, X. Che, and J. He, "Few-shot fine-grained image classification: A survey," in *ICNLP*, 2022, pp. 201–211.

[16] 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 *AAAI*, 2023, pp. 2821–2829.

[17] L.-J. Zhao, Z.-D. Chen, Z.-X. Ma, X. Luo, and X.-S. Xu, "Angular isotonic loss guided multi-layer integration for few-shot fine-grained image classification," *IEEE Trans. Image Process.*, vol. 33, pp. 3778–3792, 2024.

[18] S. Lee, W. Moon, and J.-P. Heo, "Task discrepancy maximization for fine-grained few-shot classification," in *CVPR*, 2022, pp. 5321–5330.

[19] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Tech. Rep., California Institute of Technology, 2011.

[20] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshops*, 2011.

[21] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proc. ICCV Workshops*, 2013, pp. 554–561.

[22] Y. Zhu, C. Liu, and S. Jiang, "Multi-attention meta learning for few-shot fine-grained image recognition," in *IJCAI*, 2021.

[23] 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 *AAAI*, 2024, pp. 4136–4144.

[24] Z. Zha, H. Tang, Y. Sun, and J. Tang, "Boosting few-shot fine-grained recognition with background suppression and foreground alignment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3947–3961, 2023.

[25] 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 *ACM MM*, 2021, pp. 107–115.

[26] X. Li, Z. Guo, R. Zhu, Z. Ma, J. Guo, and J.-H. Xue, "A simple scheme to amplify inter-class discrepancy for improving few-shot fine-grained image classification," *Pattern Recognit.*, vol. 156, pp. 110736, 2024.

[27] S. Yang, X. Li, D. Chang, Z. Ma, and J.-H. Xue, "Channel-spatial support-query cross-attention for fine-grained few-shot image classification," in *ACM MM*, 2024, pp. 9175–9183.

[28] L. Yu, Z. Guan, W. Zhao, Y. Yang, and J. Tan, "Adaptive task-aware refining network for few-shot fine-grained image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, pp. 2301–2314, 2025.

[29] S. Qiu, W. Yang, and M. Yang, "Hybrid feature collaborative reconstruction network for few-shot fine-grained image classification," in *ICASSP*, 2025, pp. 1–5.