International Journal for Housing Science and Its Applications

Publish September 9, 2025. Volume 47, Issue 1 Pages 783-794

https://doi.org/10.70517/ijhsa47168

CARE-Net: A Fine-Grained Image Classification Network via Cross-Layer Semantic Recalibration

Yao Yao[1,*] and Erhao Chen[2]

[1] Dundee International Institute of Central South University, Central South University, Changsha, Hunan, 410083, China

[2] School of Information Science and Technology, Hangzhou Normal University, Hangzhou, Zhejiang, 311121, China

Corresponding authors: (e-mail: 2542737@dundee.ac.uk).

Abstract Fine-grained image classification (FGIC) is increasingly important in computer vision, driven by its extensive use across various domains. However, existing methods often struggle to achieve both discriminative feature representation and semantic consistency, primarily due to subtle inter-class differences, complex object structures, and background clutter. To tackle these issues, this paper proposes a innovative framework named CARE-Net (Cross-level Adaptive Recalibration and Enhancement Network), which enhances feature learning through three synergistic mechanisms: multi-scale fusion, cross-layer guidance, and explicit feature reconstruction. Specifically, CARE-Net extracts multi-scale features from different semantic levels and employs a guidance enhancement module to recalibrate shallow features using high-level semantic cues. A lightweight attention-based module is then introduced to adaptively fuse features across scales, reinforcing responses in key discriminative regions. Finally, an auxiliary reconstruction branch is incorporated to enforce structural consistency across semantic layers under supervision. Experimental results on the CUB-200-2011 and Stanford Dogs datasets show that CARE-Net achieves Top-1 classification accuracies of 76.3% and 75.8%, respectively, outperforming several mainstream baselines. Ablation experiments provide further evidence for the effectiveness and complementary nature of each module. These results demonstrate that CARE-Net provides an efficient and interpretable solution for FGIC in complex visual environments.

Index Terms Fine-grained Image Classification, Multi-scale Fusion, Feature Reconstruction, Semantic Guidance.

I. Introduction

Fine-grained image classification focus on recognizing subcategories that belong to the same coarse-grained class but exhibit subtle visual differences. Compared to conventional image classification, FGIC typically involves objects with highly similar structures or textures, such as various species of birds[1], dogs[2], insects[3], and vehicles[4]. The key challenge lies in extracting discriminative features from minor local variations. As illustrated in Figure 1, FGIC faces two major difficulties. On the one hand, inter-class differences are minimal, with many subcategories differing only slightly in color, shape, or texture, making them difficult to differentiate. On the other hand, intra-class variations are large, as images within the same subcategory may vary significantly due to differences in viewpoint, occlusion, and complex backgrounds. This "low inter-class variance and high intra-class diversity" challenges global representation methods, demanding stronger perception of fine-grained local features.

Figure 1: Low inter-class difference and high intra-class variation

Despite the unprecedented success of CNNs in general image classification tasks, several key challenges persist in fine-grained image classification (FGIC). One significant issue is that many models fail to establish effective semantic guidance from high-level features to lower layers, resulting in shallow features struggling to focus on discriminative regions. This shortcoming undermines the task-oriented specificity and discriminative quality of the overall feature representation. To address this, several strategies have been explored, including cross-layer consistency regularization[5], [6], multi-stage guidance training[7], [8], [9], and multi-branch collaborative fusion[10]-[13]. Although these methods have achieved progress at different levels, they often suffer from limitations such as simplistic fusion strategies, weakly coupled guidance mechanisms, or insufficient modeling of contributions from heterogeneous semantic layers. These limitations hinder the effective integration of multi-scale information and restrict the model's capacity to improve its discriminative performance through cross-layer collaboration.

In addition, conventional multi-scale fusion strategies often rely on feature concatenation[14], [15]or pyramid structures[16] to integrate information across scales. These methods aim to enhance the robustness and granularity of visual representations by combining features from different receptive fields. However, when semantic discrepancies and relative contributions between scales are not adequately considered, direct fusion may lead to redundancy or semantic conflicts, ultimately degrading classification performance. On the one hand, fixed fusion strategies lack adaptive control, making it difficult to exploit the most relevant scale-specific information. On the other hand, approaches that depend solely on shallow or single-scale salient regions often overlook the guiding role of high-level semantics, resulting in incomplete or inconsistent target representations.

Finally, most existing methods rely on classification output as the sole supervisory signal, which limits the enforcement of semantic consistency across layers. Here, structural integrity refers to the preservation of spatial arrangements and local details of object parts across different feature layers, while semantic consistency denotes the alignment of high-level semantic meanings across multi-scale or cross-layer feature representations, ensuring that different layers focus on the same category-relevant regions.

To overcome the challenges mentioned above, this paper introduces a novel multi-scale enhancement framework for fine-grained image classification, termed CARE-Net. It is designed to tackle three core issues in FGIC: local discriminative representation enhancement, cross-layer semantic guidance, and structural consistency reconstruction. By constructing a unified semantic guidance mechanism and an adaptive fusion strategy, the proposed framework enables effective collaborative representation across multiple feature scales.

Built upon ResNet-50 as the backbone, CARE-Net first extracts multi-scale features from different semantic layers. A cross-layer guidance enhancement module is then introduced to explicitly recalibrate low-level detail features using high-level semantic information, thereby enhancing the model's focus on discriminative regions. Subsequently, a scale-adaptive fusion module is employed, which leverages a lightweight attention mechanism to dynamically learn the importance of each scale and perform precise feature fusion. To further improve consistency, an explicit reconstruction branch is incorporated to reconstruct semantic features at each layer and enforce a reconstruction loss, thereby aligning structural representations across scales and enhancing the model's overall discriminative capacity. The main contributions of this paper are summarized as follows:

(1) A unified network architecture is proposed to jointly model multi-scale enhancement and cross-layer collaboration, significantly improving both discriminative capability and feature consistency in FGIC.

(2) A lightweight cross-layer guidance module, termed GuidedBoost, is designed to explicitly recalibrate shallow features using high-level semantics, thereby enhancing the response to key discriminative regions.

(3) A scale-adaptive fusion mechanism is introduced, which dynamically adjusts the contributions of different semantic layers through a lightweight attention-based strategy.

(4) An explicit structural reconstruction branch is incorporated, which enforces multi-level feature consistency via reconstruction loss, improving structural awareness and robustness.

(5) Comprehensive experiments on two benchmark datasets highlight the superior performance of the proposed model, confirming its ability to improve both discriminative representation and structural consistency.

II. Related work

II. A. Fine-grained image classification methods

Existing FGIC methods are typically categorized based on the level of supervision into two groups: strongly supervised and weakly supervised approaches. Strongly supervised methods (e.g., Part-based R-CNN[17], PS-CNN[18]) rely on manually annotated fine-grained labels, such as part locations, bounding boxes, or key points. Their core idea is to leverage explicit structural cues to guide the model's attention toward discriminative regions. Although these methods achieve high classification accuracy, their reliance on detailed annotations significantly limits their scalability and applicability in cross-domain or large-scale scenarios. In contrast, weakly supervised methods depend solely on image-level labels, eliminating the need for costly part annotations or localization information. Current mainstream research in this area falls into two main categories: region proposal methods, which localize discriminative regions under weak supervision [19], and attention-based

methods, which learn to highlight informative regions without explicit annotations [20], [21]. Compared to strongly supervised approaches, weakly supervised strategies offer greater scalability and practicality, as they can guide the model to attend to key regions using only coarse-grained labels. Consequently, they have emerged as an increasingly prominent topic in recent FGIC research.

II. B. Multi-scale feature fusion mechanisms

Multi-scale feature fusion is a fundamental component in improving the effectiveness of fine-grained image recognition. Features extracted from different semantic layers carry distinct representational capacities; thus, effectively integrating these heterogeneous features is essential for enhancing the model's fine-grained discriminative ability. A representative early approach is the FPN [22], which introduces a top-down pathway structure in object detection to achieve hierarchical compensation and fusion across feature scales. This concept was later adapted to fine-grained classification tasks to enable joint modeling of local details and global semantics. For instance, Li et al. [23] proposed a spatially aligned feature pyramid network that enhances inter-class discriminability through spatial calibration and contrastive learning. Other works, such as TASN [24] and NTS-Net [25], integrate attention or region selection to strengthen multi-scale interactions. From a design perspective, fusion methods can be grouped into static structures and dynamic mechanisms. Static methods like FPN and PAFPN [26] follow predefined paths for feature integration, offering simplicity but limited flexibility. Dynamic mechanisms use attention modules to adaptively weight features based on context, providing better robustness for subtle categories. Bidirectional designs such as BiFPN [27] further improve interaction by enabling both top-down and bottom-up information flow.

Inspired by these studies, this paper proposes a scale-adaptive fusion module that jointly leverages global multi-scale features and high-level semantic cues. In contrast to static pyramids or existing dynamic attention mechanisms, the proposed module dynamically models the importance of each scale conditioned on semantic relevance, rather than relying solely on saliency. This semantic-conditioned weighting reduces redundancy from irrelevant scales, alleviates semantic conflicts, and yields a unified, more discriminative feature representation.

II. C. Cross-layer semantic guidance mechanisms

Although multi-scale feature fusion enables the integration of information across different semantic levels, a significant semantic gap often exists between shallow and deep features in practical applications. Shallow features typically lack semantic abstraction, while deep features lose fine-grained details. Directly fusing them can introduce redundant information or semantic conflicts, thereby limiting the model's discriminative power. To address this issue, recent studies have explored cross-layer semantic guidance mechanisms. The core idea is to use high-level semantic features to guide the alignment of shallow structural representations, enabling low-level features to focus more effectively on category-relevant regions. For instance, API-Net [28] constructs bidirectional feature vectors to model latent semantic differences between image pairs and generates semantic gating vectors for image-level contrastive guidance. SCANet [29] introduces the HSF module, which explicitly aligns shallow spatial details by leveraging deep semantic features through attention-based guidance, thus narrowing the semantic gap between feature levels.

Building upon this line of research, this paper introduces a cross-layer semantic guidance mechanism on top of multi-scale feature extraction. High-level semantic features are used as guidance signals and are explicitly propagated to multiple shallow feature branches before fusion. This process modulates channel-wise attention responses, enabling shallow features to focus more precisely on semantically relevant regions. Unlike prior guidance approaches that typically inject semantic cues into limited layers or treat guidance and fusion as separate processes, our method applies lightweight channel-wise guidance uniformly to all shallow branches prior to fusion and directly couples it with the scale-adaptive fusion module. This tight integration enhances semantic consistency and complementarity between feature layers, leading to more discriminative and robust fused representations.

II. D. Explicit refactoring strategies

In FGIC tasks, in addition to enhancing discriminative modeling, preserving the structural integrity of features is equally important. To enhance the model's capacity for capturing local features, explicit reconstruction strategies have been introduced as auxiliary training mechanisms. These strategies formulate reconstruction tasks—such as local structural recovery—to drive the model to maintain structural consistency in the feature space, thereby increasing sensitivity to key regions. Unlike attention mechanisms that focus on region selection, reconstruction strategies emphasize structure preservation, making the two approaches complementary. A representative example is DCL [30], which perturbs discriminative regions and guides the model to reconstruct their structure, thereby indirectly strengthening its discriminative awareness of critical areas. Furthermore, bidirectional feature reconstruction networks[31] introduce a dual-directional strategy that allows support features to reconstruct query features and vice versa, simultaneously enhancing inter-class separability and suppressing intra-class

variability. HFCR-Net [32] further exploits both channel-wise and spatial reconstruction to preserve fine-grained structural information across multiple feature scales, Strengthening the model's capacity to detect subtle distinctions.

Based on these insights, this paper introduces an explicit reconstruction mechanism as an auxiliary branch. Operating on multi-scale fused features, it leverages both shallow structure and deep semantics to perform local structural restoration. In contrast to previous reconstruction-based methods that operate on single-scale or part-level features and treat reconstruction independently of fusion, our approach reconstructs each scale's features directly from the fused representation, thereby enforcing cross-scale consistency. This joint design not only regularizes the fusion process but also preserves semantic integrity across layers, enhancing both the stability and interpretability of intermediate representations.

III. Methodology

III. A. Overall network structure

This study aims to establish a fine-grained image classification framework with integrated guidance and multi-scale consistency enhancement. The backbone adopts ResNet-50 due to its clear hierarchical architecture and strong extensibility, providing a stable semantic foundation for subsequent cross-layer guidance and multi-scale fusion. The training procedure is divided into four stages: First, the input image is processed by the backbone to extract shallow-to-deep features (layer1–layer4), denoted as v_2, v_3, v_4, and v_5. These features are passed through 1×1 convolutions to unify the channel dimensions, forming aligned multi-scale intermediate representations. Second, a cross-layer semantic guidance module is introduced. It uses the deep feature v_5 to enhance v_2, v_3, and v_4 through weighted attention, injecting more discriminative semantic information into lower-level features. Third, all features are upsampled to a uniform spatial resolution and fused using a scale-adaptive fusion module. This module employs a lightweight attention mechanism to generate adaptive fusion weights, effectively integrating features from different scales into a unified discriminative representation. Finally, the explicit reconstruction supervision module is incorporated during training. This branch applies reconstruction loss to all four guided features to improve feature fidelity and cross-scale consistency. The primary task is image classification, while the auxiliary task enforces structural reconstruction. The overall pipeline is illustrated in Figure 2.

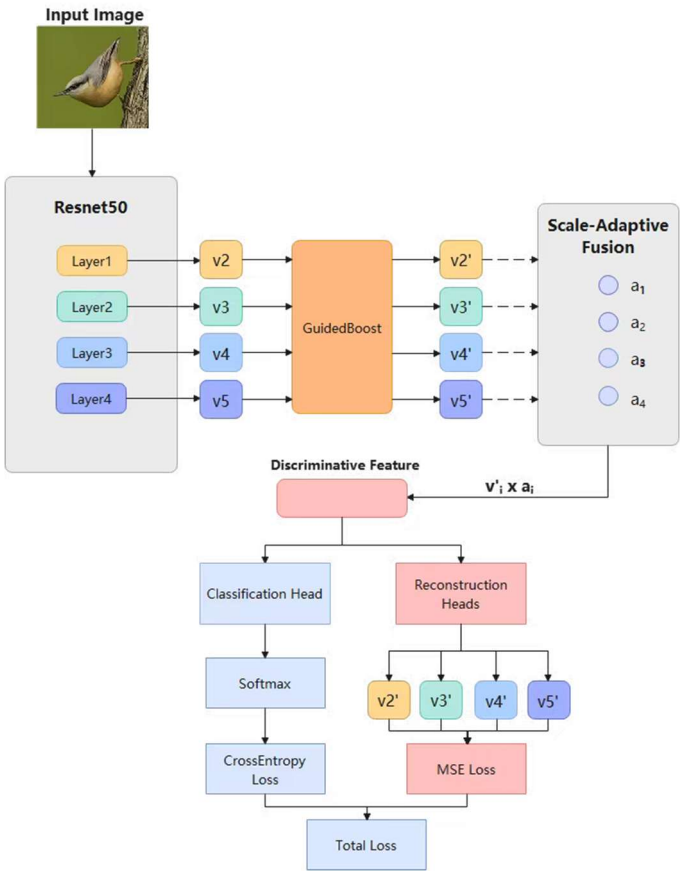

Figure 2: The network structure of CARE-Net

III. B. Cross-layer semantic guidance module

This study extracts multi-scale features from different layers of the ResNet-50 backbone, denoted as v_2 (layer1), v_3 (layer2), v_4 (layer3), and v_5 (layer4). Among them, v_2 and v_3 are obtained from shallow layers and preserve more fine-grained texture and edge information, while v_4 and v_5 are derived from deeper layers and possess stronger semantic abstraction capabilities. In fine-grained image classification tasks, although low-level features contain abundant textures and edges, they lack semantic expression ability and are easily affected by background noise, making it difficult to generate stable responses in target regions.

To enhance the discriminative ability of shallow features, considering their lack of semantic information, this paper designs a cross-layer guidance mechanism. It introduces high-level semantic features as guidance signals to enhance the discriminative features of shallow layers, thereby alleviating the problem of semantic inconsistency and insufficient focus on key regions.

Specifically, high-level feature v_5 is used to guide the low-level features v_2, v_3, and v_4. Due to spatial resolution differences among multi-scale features, the guiding feature v_5 needs to be aligned to the spatial dimensions of each lower-scale feature. First, v_5 is upsampled to the same spatial resolution as v_i using bilinear interpolation, denoted as $\hat{v}_5$, as shown in Equation (1):

$$\hat{v}_5 = BilinearUpsample\left(v_5, size = H \times W\right) \tag{1}$$

To extract the most critical semantic channels from the guidance signal and improve its discriminative precision, a lightweight channel attention module is introduced to enhance the upsampled feature $\hat{v}_5$. Specifically, this module captures global contextual information via global average pooling and applies two consecutive 1×1 convolutional layers with non-linear activations to generate channel-wise importance weights. Compared to general attention mechanisms such as SE[33] and ECA[34], the proposed structure is more lightweight and computationally efficient, making it well-suited for use as a semantic guidance generator in compact modules. The process is defined in Equation (2):

$$A_i = \sigma\left(Conv1\times1\left(ReLU\left(Conv1\times1\left(Pool\left(\hat{v}_5\right)\right)\right)\right)\right) \tag{2}$$

Here, $Pool$ represents the global average pooling, and σ indicates Sigmoid activation.

Subsequently, the attention map is applied to the target feature v_i to produce the enhanced feature v_i', as defined in Equation (3):

$$v_i^{'} = v_i \cdot \left(1 + \lambda \cdot A_i\right), i \in \{2,3,4\} \tag{3}$$

λ denotes a scaling factor that controls the strength of semantic guidance. In this study, λ is set to 0.2 to preserve the structural integrity of the target features while enabling moderate semantic injection. This parameter remains fixed during training and is excluded from gradient updates.

The enhanced multi-scale features v_2' to v_5' are subsequently passed into the scale-adaptive fusion module for further weighted integration, yielding the final discriminative feature representation. This representation provides enriched semantic support for both classification and reconstruction tasks.

III. C. Scale adaptive fusion module

Given that a single-scale representation struggles to balance fine-grained details and high-level semantics, this module introduces a scale-adaptive fusion strategy to further integrate semantic and detail information across different feature levels. The core idea is to leverage high-level semantic features to dynamically weight the responses of multi-scale features, thereby adjusting their contributions to the final discriminative representation.

Due to significant spatial resolution differences among features from different scales, direct fusion would lead to dimensional inconsistencies, hindering effective information interaction and integration. Therefore, before applying scale-aware weighting, all enhanced features (i.e., v_i' obtained in Section 3.2) are spatially aligned. To simplify computation and unify dimensions, the resolution of the lowest-level feature v_2' is used as the reference. Other features are upsampled via bilinear interpolation, as formulated in Equation (4), to ensure consistent spatial dimensions and enable pixel-wise fusion in a shared spatial domain:

$$\hat{v}_i = BilinearUpsample\left(v_i, size = H \times W\right) \tag{4}$$

Subsequently, for each enhanced feature v_i', a global semantic vector g_i is extracted via global average pooling. These vectors are then concatenated to form a feature set $G = \left[g_2; g_3; g_4; g_5\right] \in \mathrm{R}^{4\times C}$. The aggregated vector $\bar{g}$, representing the overall distribution trend across scales, is computed by averaging the pooled vectors as in Equation (5). It is then concatenated with the high-level semantic vector g_5, and the result is passed through a two-layer MLP followed by SoftMax function to obtain the scale fusion weights α, as shown in Equation (6):

$$\bar{g} = \frac{1}{4}\sum_{i=2}^{5} g_i \tag{5}$$

$$\alpha = Softmax\left(MLP\left(\left[\bar{g}; g_5\right]\right)\right) \tag{6}$$

Here, $\alpha \in \mathrm{R}^4$ denotes the fusion weights for each scale. These weights are then used to perform a weighted combination of the spatially aligned features, resulting in the final discriminative representation v_f, as described in Equation (7):

$$v_f = \sum_{i=2}^{5} \alpha_i \cdot \tilde{v}_i \tag{7}$$

III. D. Explicit reconstruction supervision module

In Section 3.3, the fused discriminative features v_f were obtained. This module further guides these enhanced features to encourage the fused representation to retain more structural details and semantic consistency, thereby improving the model's ability to perceive fine-grained differences.

First, these features are input into the reconstruction branches R_i corresponding to the four scales, producing the reconstructed features $\tilde{v}_i$ for each scale. Each reconstruction branch uses a 1×1 convolutional layer to maintain channel consistency and control computational complexity. Then, the semantic-enhanced target feature v_i' is used as a supervisory signal to minimize the mean squared error (MSE) between $\tilde{v}_i$ and v_i', guiding the discriminative features to retain multi-scale semantic information. The computation is described in Equation (8):

$$\mathbf{L}_{recon} = \sum \boldsymbol{i} = 2^5 \left\| \tilde{\boldsymbol{v}}_{\boldsymbol{i}} - \boldsymbol{v}_{\boldsymbol{i}}' \right\|_2^2 \tag{8}$$

This reconstruction process not only serves as an auxiliary supervision signal to effectively enhance the semantic integrity of the discriminative features but also mitigates the fusion conflicts between shallow texture information and deep semantic features. By forcing the discriminative features to possess cross-scale reconstruction capability, the model learns more discriminative and interpretable intermediate representations.

The final training objective consists of both the classification loss and the reconstruction loss, with the overall loss function defined as:

$$\mathrm{L}_{total} = \mathrm{L}_{cls} + \lambda_{recon} \cdot \mathrm{L}_{recon} \tag{9}$$

where $\mathbf{L}_{cls}$ denotes the cross-entropy classification loss, and λ_{recon} is a hyperparameter that controls the influence of the reconstruction loss on the total loss.

This explicit reconstruction mechanism, as an auxiliary task, significantly improves the stability and structural reversibility of the discriminative features during training. It helps alleviate the overfitting of the backbone network to category labels, ultimately enhancing the model's generalization and discriminative capability in fine-grained image classification tasks.

IV. Experimental design and result analysis

IV. A. Experimental setup

IV. A. 1) Datasets

Two fine-grained image classification datasets are employed in this study: CUB-200-2011 and Stanford Dogs. The former contains 200 bird subcategories, while the latter covers 120 dog breeds. Both datasets provide a large number of images with detailed annotations and are widely used for evaluating the performance of fine-grained classification models. Detailed statistics are presented in Table 1.

Table 1: Datasets details

Dataset	Category	Training	Testing	Total
CUB-200-2011	200	5994	5794	11788
Stanford Dogs	120	12000	8580	20580

IV. A. 2) Evaluating indicator

To comprehensively evaluate the performance of CARE-Net on FGIC tasks, Top-1 classification accuracy and model parameter size are adopted as the primary evaluation metrics.

Top-1 accuracy measures the proportion of samples in the test set for which the predicted label matches the ground truth. As one of the most commonly used and intuitive metrics, it directly reflects the model's discriminative ability across fine-grained categories. The calculation is defined in Equation (10):

$$Acc = \frac{N_{correct}}{N_{total}} \tag{10}$$

where $N_{correct}$ denotes the number of correctly predicted samples, and N_{total} represents the total number of samples in the test set.

Model parameter size reflects the structural complexity and overall scale of the model. This metric impacts not only the training and inference efficiency but also the model's deploy ability in resource-constrained environments. For typical deep neural networks, the total number of parameters is the sum of all learnable parameters across layers. The parameter counts for convolutional and fully connected layers are calculated as follows:

For convolutional layers:

$$ParamsConv = K_h \times K_w \times C_{in} \times C_{out} + C_{out} \tag{11}$$

where K_h and K_w denote the height and width of the convolution kernel, and C_{in}, C_{out} are the numbers of input and output channels, respectively.

For fully connected layers:

$$ParamsFC = in_{features} \times out_{features} + out_{features} \tag{12}$$

IV. A. 3) Experimental details

All experiments were conducted using PyTorch 2.5.1 on Ubuntu 22.04 with Python 3.12. The training was performed on a single NVIDIA RTX 4090D GPU, paired with a 16-core Intel Xeon Platinum 8481C CPU. CUDA 12.4 was utilized to accelerate computation, ensuring efficient and stable training. The backbone network adopted is ResNet50, initialized with ImageNet pre-trained weights to accelerate convergence and enhance feature representation capability. All comparison methods were trained with identical data preprocessing and training strategies to ensure fair evaluation.

Input images were resized to 224×224, and standard data augmentation techniques were applied during training, including random cropping and random horizontal flipping, to improve generalization. During testing, scaling and center cropping were used for consistent evaluation. All images were normalized using ImageNet statistics: mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225].

Training was performed using the SGD optimizer with an initial learning rate of 0.005, momentum of 0.9, and weight decay of 1e-4. The batch size was set to 16, and training was conducted for a total of 60 epochs. At the end of each epoch, the model was evaluated on the validation set. This configuration was determined through preliminary tuning and demonstrated good convergence speed and stability, making it well-suited for small-batch, fine-grained classification tasks.

When the feature reconstruction module was enabled, MSE loss was introduced as an auxiliary objective. The loss weight coefficient λ was set to 0.3, a value empirically selected to ensure training stability while significantly improving classification accuracy. The classification task was optimized using the cross-entropy loss, and the model was trained by jointly optimizing both classification and reconstruction losses, achieving a synergistic enhancement of multi-scale feature fusion and attention guidance.

IV. B. Comparative experiment

Comparison experiments were conducted on two widely used benchmark datasets: CUB-200-2011 and Stanford Dogs. The compared models cover a broad spectrum of mainstream approaches, including traditional convolutional networks (e.g., ResNet18, ResNet50, DenseNet201), recently proposed lightweight architectures (e.g., GhostNetV2_160, EfficientNet_b4), and high-performance representative models (e.g., Inception_v3, Swin-Tiny, DeiT-S), thereby encompassing diverse model scales and architectural paradigms.

As shown in the experimental results, CARE-Net achieves competitive performance on both datasets, reaching 76.6% accuracy on CUB-200 and 75.8% on Stanford Dogs, second only to DeiT-S. Compared to the ResNet50 baseline, CARE-Net improves classification accuracy by 1.3% and 1.2%, respectively, demonstrating the generalizability and stability of the proposed semantic enhancement mechanism in fine-grained recognition tasks.

It is noteworthy that although CARE-Net does not currently achieve the highest absolute accuracy, it delivers results that are very close to the top-performing DeiT-S (77.5% vs. 76.6% on CUB-200, 76.7% vs. 75.8% on Stanford Dogs) while significantly outperforming Swin-Tiny by large margins (10.3% and 12.7% absolute gains, respectively). This demonstrates that the proposed semantic enhancement mechanism is highly effective even without relying on advanced transformer backbones. Nevertheless, the slight performance gap compared to DeiT-S may be largely attributed to the relatively outdated

backbone of CARE-Net (ResNet50), which limits its ability to capture long-range dependencies and fine-grained global context to the same extent as modern vision transformers.

In summary, CARE-Net achieves near-optimal performance while maintaining a compact model size, showcasing a favorable trade-off between accuracy and computational cost. A promising direction for future improvement is to integrate the proposed semantic enhancement framework with more advanced backbones—such as DeiT—potentially enabling CARE-Net to surpass current state-of-the-art results while retaining its efficient multi-scale fusion and reconstruction modules.

Table 2: Comparative experiment on CUB-200-2011 and Stanford Dogs

Model	Params (M)	CUB-200 Acc (%)	Stanford Dogs Acc (%)
Efficientnet_b4	19.3	76.0	75.4
Inception_v3	23.8	75.4	74.7
Densenet201	20.0	76.1	75.3
Swin-Tiny	28.3	66.3	63.1
DeiT-S	22.0	**77.5**	**76.7**
GhostnetV2_160	12.4	76.5	75.5
Resnet18	11.7	72.8	70.7
Resnet50	25.6	75.3	74.6
Care-net	25.0	**76.6**	**75.8**

IV. C. Ablation experiment

IV. C. 1) Key module ablation experiment

To measure the contribution of each essential module in CARE-Net, ablation experiments were conducted on the CUB-200-2011 dataset. Table 3 presents the results of the comparative study, demonstrating that the inclusion of each module positively contributes to the model's overall classification performance. In Experiment 1, the model was trained using only the standard ResNet50 backbone without any structural enhancements, achieving a Top-1 accuracy of 75.3%. In Experiment 2, the multi-scale fusion module was added, resulting in a 0.3% accuracy improvement, indicating that collaborative modeling across feature layers helps capture fine-grained differences. Experiment 3 further introduced the semantic guidance mechanism, which utilizes high-level features to regulate and enhance lower-level representations. This led to an additional 0.4% performance gain compared to Experiment 2, validating the effectiveness of semantic attention in modeling discriminative regions. In Experiment 4, the feature reconstruction branch was incorporated, completing the full CARE-Net architecture. The model achieved an accuracy of 76.6%, marking a 1.3% improvement over the baseline in Experiment 1. The reconstruction module imposes explicit constraints across different semantic layers, enhancing feature consistency and robustness. It also helps alleviate overfitting, delivering performance gains without significantly increasing model complexity.

These results confirm the complementary synergy among the three proposed modules and demonstrate the cumulative performance improvements brought by progressive integration. This provides empirical evidence supporting the modular design of the network and offers practical guidance for future architectural extensions.

Table 3: Ablation experiment of Care-net on CUB-200-2011

Experiment	Multi-scale fusion	Semantic guidance	Reconstruction Branch	Accuracy (%)
1	×	×	×	75.3
2	√	×	×	75.6
3	√	√	×	75.9
4	√	√	√	76.6

IV. C. 2) Effect of the Reconstruction Branch and the Weight Coefficient λ

This section delves deeper into the effect of the reconstruction branch on the model's performance by analyzing the effect of varying the reconstruction loss weight coefficient λ on classification accuracy. As shown in Table 4, increasing λ initially improves performance, followed by a decline, with the best accuracy observed at $\lambda = 0.3$. This suggests that assigning approximately 20–30% of the total loss to the reconstruction term facilitates more effective modeling of structural consistency. When λ is too small, the reconstruction loss contributes minimally to the overall optimization, limiting the network's ability to leverage the structural information induced by the auxiliary reconstruction task, thereby resulting in marginal performance gains. Conversely, a large λ leads to an overemphasis on the reconstruction objective, causing the model to focus excessively on low-level detail recovery during training. This imbalance compromises the learning of discriminative high-level semantic

features necessary for classification. Such a shift in optimization priorities results in a misalignment between the learned representations and the primary classification objective, ultimately degrading the model's accuracy.

Table 4: Reconstruction loss weight experiment on CUB-200-2011

λ	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Acc/%	76.05	76.13	76.29	76.37	76.64	76.28	75.52	74.89	74.71

IV. D. Visualization experiment

IV. D. 1) GradCAM visualization

To intuitively demonstrate the effectiveness of the proposed method, Grad-CAM is employed for comparative visualization. Attention heatmaps generated by the backbone network ResNet50 and the proposed CARE-Net are presented for multiple category samples. As illustrated in Figure 3, the left side of each image pair shows the heatmap produced by ResNet50, while the right side displays that of CARE-Net. Overall, CARE-Net exhibits significantly improved attention localization compared to ResNet50, consistently focusing on semantically meaningful and discriminative regions across most samples. For example, in Class1 and Class2, ResNet50 produces relatively dispersed attention maps, often extending into background or irrelevant areas, indicating semantic drift. In contrast, CARE-Net accurately concentrates on key semantic parts such as the bird's head and body, demonstrating stronger focus and discriminative power. In Class3, CARE-Net's attention is more compact and less affected by background noise, reflecting enhanced feature representation capability. In challenging cases like Class4, where complex backgrounds and occlusions are present, ResNet50 tends to be distracted by irrelevant regions, leading to attention misalignment. CARE-Net, however, consistently attends to the bird's facial area, showcasing greater robustness and superior localization of semantic regions.

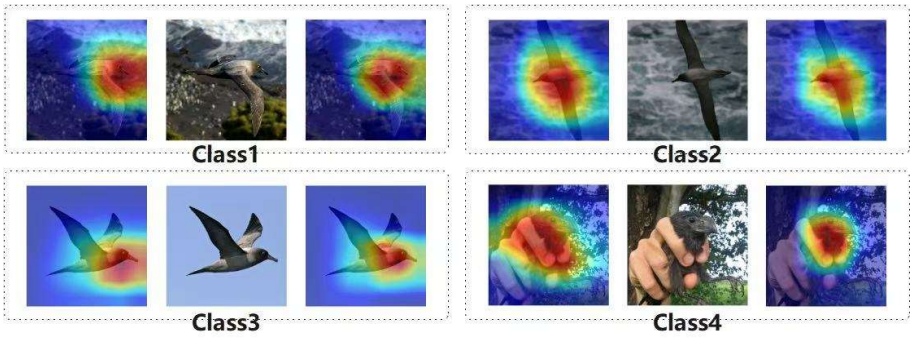

Figure 3: Grad-CAM visualization comparison

IV. D. 2) T-SNE visualization

To further analyze the discriminative capacity of the model at the feature representation level, t-SNE is employed to visualize the embedded discriminative features extracted by CARE-Net and ResNet50. The results are shown in Figure 4 and Figure 5, where each point represents a test image in the feature space, and different colors and shapes denote different classes.

As observed from the visualization, CARE-Net demonstrates superior intra-class compactness and inter-class separability in the feature space. Most categories form tighter and more distinct clusters in the two-dimensional space. For instance, Class 1, Class 29, Class 43, and Class 183 exhibit clear clustering boundaries, indicating that the features extracted by CARE-Net are more discriminative and structurally coherent. In contrast, ResNet50 shows more noticeable class overlap and intra-class dispersion in certain regions, particularly around Class 43, Class 113, and Class 155, where feature points exhibit significant overlap and spread. This suggests that ResNet50's feature representations lack sufficient class-wise discriminability, making it difficult to achieve effective clustering in the low-dimensional embedding space—largely due to the absence of explicit semantic alignment mechanisms. In summary, the t-SNE visualization provides additional evidence that CARE-Net achieves better modeling of class boundaries and semantic clusters at the feature level, highlighting its stronger feature representation capability and generalization potential in fine-grained image classification tasks.

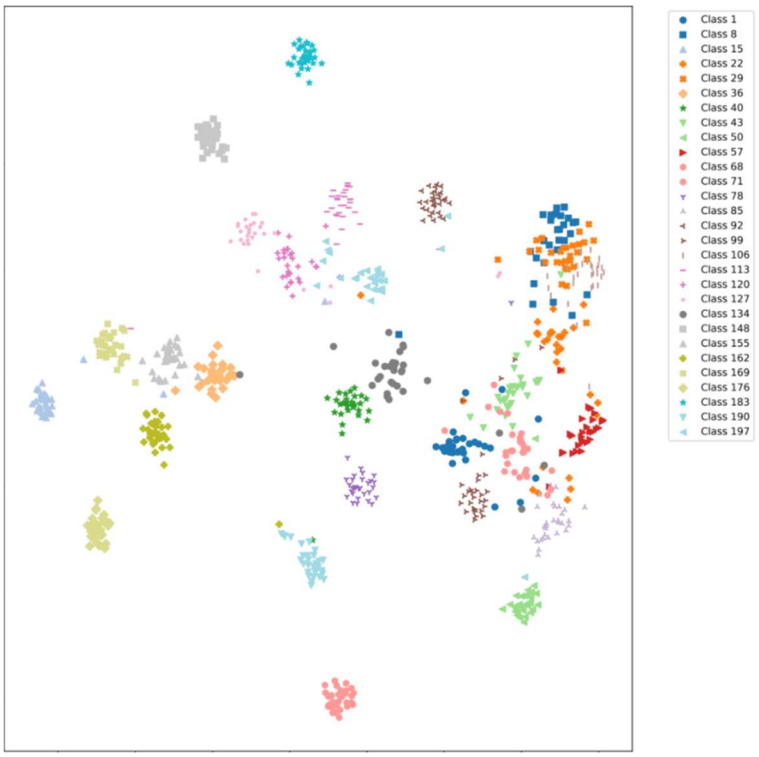

Figure 4: Care-net t-SNE visualization figure

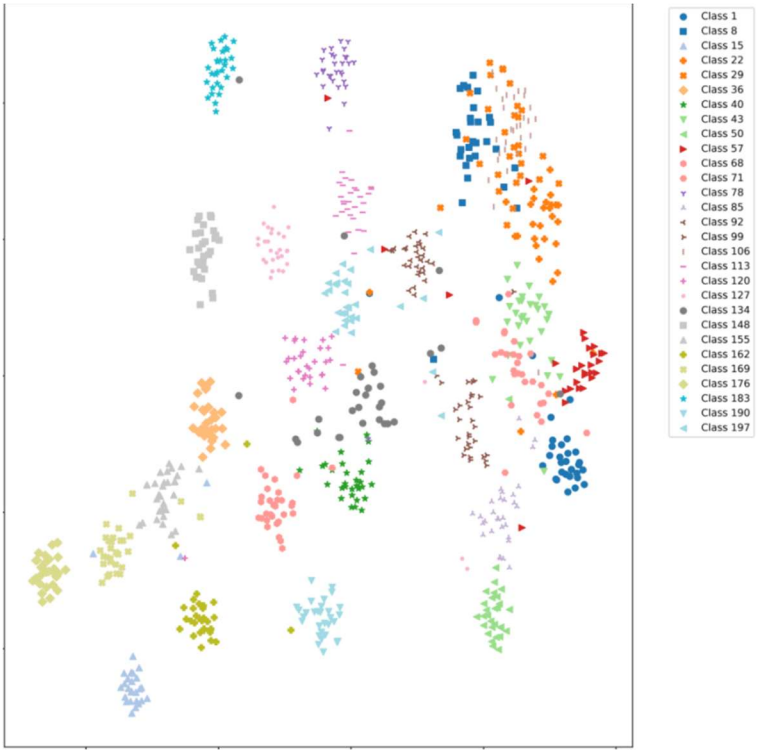

Figure 5: Resnet50 t-SNE visualization figure

V. Conclusions

This paper proposes an innovative framework for FGIC task, termed CARE-Net, which enhances the model's ability to perceive fine-grained variations by integrating three key components: a cross-layer semantic guidance module, a scale-adaptive fusion mechanism, and an explicit reconstruction supervision branch. In CARE-Net, the cross-layer guidance module

explicitly enhances shallow features using high-level semantic information, enabling the model to more precisely focus on category-discriminative local regions. The scale-adaptive fusion module dynamically integrates multi-scale discriminative information through a lightweight attention mechanism, producing structurally consistent and semantically unified feature representations. The explicit reconstruction branch introduces auxiliary structural constraints to maintain semantic consistency and structural completeness across different feature scales.

Extensive experiments demonstrate that CARE-Net achieves competitive performance on multiple fine-grained benchmarks, while maintaining moderate model complexity, thereby exhibiting strong stability and practical applicability.

Despite achieving a favorable balance between performance and complexity, CARE-Net currently adopts ResNet-50 as its backbone, which limits its capacity for deeper semantic modeling. In future work, the framework will be extended by integrating more advanced backbone networks, such as Swin Transformer, to further exploit high-level semantic representations. Moreover, additional directions will be explored, including model lightweighting and cross-modal semantic guidance, to enhance the adaptability and generalizability of CARE-Net in more complex real-world scenarios.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data sharing agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

[1] Pang W, Song W. Cross-Granularity Fusion Network for Fine-Grained Image Classification[C]//Proceedings of the 2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). Nanjing, China: IEEE, 2023: 617–622.

[2] Wang P, Gu Y, Tang Z, Zhou F. Fine-Grained Image Classification Based on Self-Distillation and RFCAConv[C]//Proceedings of the 2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS). Yanji, China: IEEE, 2024: 378–383.

[3] Mei A, Huo H. Abnormal Clustering and Cross Slicing Transformer for Insect Fine-Grained Image Classification[C]//Proceedings of the 2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV). Zhenjiang, China: IEEE, 2024: 135–141.

[4] Zhao X, Chen C. Multi-Scale Localization and Attention Mechanism for Fine-Grained Image Classification[C]//Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). Shenzhen, China: IEEE, 2024: 1173–1178.

[5] Ye Z, Hu F, Liu Y, Xia Z, Lyu F, Liu P. Associating Multi-Scale Receptive Fields for Fine-Grained Recognition[C]//Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, UAE: IEEE, 2020: 1851–1855.

[6] Luo W, et al. Cross-X Learning for Fine-Grained Visual Categorization[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 8241–8250.

[7] Chang D, Zheng Y, Ma Z, Du R, Liang K. Fine-Grained Visual Classification via Simultaneously Learning of Multi-regional Multi-grained Features [EB/OL]. arXiv:2102.00367, 2021.

[8] Eshratifar A E, Eigen D, Gormish M, Pedram M. Coarse2Fine: A Two-stage Training Method for Fine-grained Visual Classification [EB/OL]. arXiv:1909.02680, 2019.

[9] Du R, Chang D, Bhunia AK, Xie J, Ma Z, Song Y-Z, Guo J. Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches[C]//European Conference on Computer Vision (ECCV), 2020: 123 – 139.

[10] Zhang F, Li M, Zhai G, Liu Y. Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization [C]// Proc. of MultiMedia Modeling (MMM). Cham: Springer, 2021: 146–158.

[11] Shang Y, Huo H. A study on fine-grained image classification algorithm based on ECA-NET and multi-granularity [J]. International Journal of Frontiers in Engineering Technology, 2023, 5(2): 31–38.

[12] Qin W, Lu T, Zhang L, Peng S, Wan D. Multi-Branch Deepfake Detection Algorithm Based on Fine-Grained Features [J]. Computers, Materials & Continua, 2023, 77(1): 467–490.

[13] Fan Z, Li M, Zhai G, Liu Y. Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization [EB/OL]. arXiv:2003.09150, 2020.

[14] Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 4438–4446.

[15] Zheng H, Fu J, Zha Z J, et al. Learning deep bilinear transformation for fine-grained image representation[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 3846–3855.

[16] Liu S, Qi L, Qin H, et al. Path Aggregation Network for Instance Segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 8759–8768.

[17] Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for Fine-grained Category Detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014: 834–849.

[18] Huang S, Xu Z, Tao D, Zhang Y. Part-Stacked CNN for Fine-Grained Visual Categorization [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 1173–1182.

[19] Wang R, Zou W, Pu J, Wang J. Weakly Supervised Localization with Patch Detector for Fine-Grained Image Retrieval[C]//Proceedings of the 2021 IEEE International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT). Qingdao, China: IEEE, 2021: 777–780.
[20] Chen J. Weakly Supervised Learning of Discriminative Features for Fine-Grained Visual Categorization[C]//Proceedings of the 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). Fuzhou, China: IEEE, 2020: 176–180.
[21] He X, Peng Y, Zhao J. Fast Fine-Grained Image Classification via Weakly Supervised Discriminative Localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(5): 1394–1407.
[22] Lin T Y, Dollár P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2117–2125.
[23] Li Yifan, Chen Liang, Li Wei. Fine-Grained Ship Recognition With Spatial-Aligned Feature Pyramid Network and Adaptive Prototypical Contrastive Learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1–13.
[24] Zheng Haoyu, Fu Jianlong, Zha Zhengjun, Luo Jie. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 5007–5016.
[25] Yang Z, Luo T, Wang D, Hu Z, Gao J, Wang L. Learning to Navigate for Fine-grained Classification[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 438-454.
[26] Liu S, Qi L, Qin H, Shi J, Jia J. Path Aggregation Network for Instance Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018: 8759–8768.
[27] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 10781–10790.
[28] Wang S, Wang Z, Li H, et al. Semantic-Guided Information Alignment Network for Fine-Grained Image Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6558–6570.
[29] Lu G, Yao S, Li Y, et al. A Semantic-Guided Cross-Attention Network for Change Detection in High-Resolution Remote Sensing Images[J]. Remote Sensing, 2025, 17(10): 1749.
[30] Chen Y, Bai Y, Zhang W, Mei T. Destruction and Construction Learning for Fine-Grained Image Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5152–5161.
[31] WU J J, CHANG D L, SAIN A, et al. Bi-directional Feature Reconstruction Network for Fine-Grained Few-Shot Image Classification [EB/OL]. arXiv:2211.17161, 2022.
[32] Qiu S, Yang W, Yang M. Hybrid Feature Collaborative Reconstruction Network for Few-Shot Fine-Grained Image Classification [EB/OL]. arXiv:2407.02123, 2024.
[33] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 7132–7141.
[34] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 11534–11542.