

QGF-Net: Fine-Grained Image Classification via Self-Supervised Vision Transformer and Quantum Measurement Attention

xueliang Song

Chongqing Normal University

Yumin Dong

dym@cqnu.edu.cn

Chongqing Normal University

Shuang Wu

Chongqing Normal University

Bo Wang

North University of China

Research Article

Keywords: fine-grained image classification, self-supervised learning, vision transformer, quantum measurement attention, hybrid quantum-classical learning

Posted Date: August 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9637288/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

QGF-Net: Fine-Grained Image Classification via Self-Supervised Vision Transformer and Quantum Measurement Attention

Xueliang Song², Yumin Dong^{1*}, Shang Wu², Bo Wang³

^{1*}College of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China.

²College of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China.

³School of Software, North University of China, Taiyuan, Shanxi, 030051, China.

*Corresponding author(s). E-mail(s): dym@cqnu.edu.cn;

Contributing authors: flin6015@gmail.com; 910998382@qq.com;
13700582817@163.com;

Abstract

Fine-grained image classification (FGIC) aims to distinguish subordinate categories within the same basic-level class, and remains highly challenging due to small inter-class differences, large intra-class variations, and the fact that discriminative cues are often concentrated in local regions [2, 3]. Recently, self-supervised Vision Transformer (ViT) models have achieved remarkable success in visual representation learning and provide powerful global semantic features for downstream classification tasks [10, 11]. However, existing methods still suffer from two limitations. First, self-supervised ViT tends to focus more on global semantics while being insufficiently sensitive to subtle local discriminative regions that are crucial for fine-grained recognition [7, 8]. Second, conventional attention mechanisms mainly rely on classical similarity computation, which may be insufficient for modeling complex high-order nonlinear dependencies among local regions; meanwhile, recent quantum-enhanced vision studies suggest that parameterized quantum circuits can provide an alternative mechanism for modeling structured dependencies [19, 20].

To address these issues, this paper proposes QGF-Net, a hybrid quantum-classical fine-grained image classification network that integrates self-supervised visual representations, illumination-consistent local feature stabilization, local

discriminative region proposal, and quantum measurement attention into a unified architecture. Specifically, a DINOv2-pretrained ViT is first employed to extract global semantic features and patch token representations from input images [11]. Then, an Illumination-Consistent Local Coarse Fusion (ICLCF) module is introduced to improve the robustness and consistency of local token features under illumination changes. Based on the stabilized local representations, a Local Discriminative Region Proposal (LDRP) module is designed to identify the most informative local regions, thereby narrowing the search space for fine-grained recognition. A Quantum Measurement Attention (QMA) module is further constructed to model the relationships among key local regions through a lightweight parameterized quantum circuit, where qubit measurement results are used to generate adaptive attention weights for enhancing local discriminative representations [18–20]. Finally, the global semantic features and quantum-enhanced local features are fused and fed into a Hybrid Quantum-Classical Classifier (HQCC) for final prediction.

Experimental results on three representative fine-grained benchmarks, namely CUB-200-2011, Stanford Cars, and FGVC-Aircraft, demonstrate that QGF-Net consistently outperforms strong CNN, Transformer, and locally enhanced baselines in terms of Top-1 Accuracy, Macro-F1, and robustness [13, 15, 16]. The full model achieves 92.0%, 94.2%, and 89.5% Top-1 accuracy on CUB, Cars, and Aircraft, respectively, with an average Top-1 accuracy of 91.9%. Robustness experiments under brightness, shadow, and occlusion perturbations further show that QGF-Net achieves the smallest average accuracy drop among compared methods. Ablation studies verify the effectiveness of ICLCF, LDRP, QMA, and HQCC, and additional sensitivity analysis demonstrates that the proposed quantum module provides a favorable balance between representation gain and training stability. Overall, QGF-Net provides a promising paradigm for fine-grained image classification by combining self-supervised ViT with hybrid quantum-classical modeling.

Keywords: fine-grained image classification, self-supervised learning, vision transformer, quantum measurement attention, hybrid quantum-classical learning

1 Introduction

Fine-grained image classification (FGIC) aims to distinguish subordinate categories within the same basic-level class, such as bird species, car models, and aircraft variants [2, 3]. Compared with conventional image classification, FGIC is substantially more challenging because inter-class differences are often subtle, while intra-class variations caused by pose, viewpoint, illumination, background clutter, and occlusion can be significant [8, 24]. In many fine-grained scenarios, the most discriminative cues are concentrated in small local regions rather than the entire object [4, 5]. For example, in bird recognition, subtle differences in the head texture, wing pattern, or tail shape may determine the final category. Therefore, effective FGIC requires not only strong global semantic understanding but also accurate modeling of subtle local discriminative details [7, 8].

Recent advances in deep learning have significantly improved the performance of FGIC. Earlier studies mainly relied on convolutional neural networks and part-based learning strategies to localize informative regions and enhance local feature representation [4, 5]. More recently, Vision Transformers (ViTs) have demonstrated strong capability in visual recognition by modeling long-range dependencies through self-attention [10]. Meanwhile, self-supervised learning has become an effective paradigm for learning transferable visual representations from unlabeled data, alleviating the heavy dependence on large-scale manual annotations [11, 12]. Benefiting from these advances, self-supervised pretrained ViTs provide a promising foundation for fine-grained classification by offering rich semantic features and stronger generalization ability [11, 12].

Despite their strong representation power, existing self-supervised ViT-based methods still exhibit several limitations when applied to FGIC. First, self-supervised ViT models are generally optimized to capture holistic semantic consistency, which makes them more sensitive to global object-level information than to subtle local discriminative regions [11, 12]. As a result, fine-grained category-specific details may not be sufficiently emphasized in the learned representations. Second, in practical fine-grained scenarios, local discriminative regions are often affected by illumination changes, shadows, and reflections. Semantically consistent local regions may exhibit significantly different feature responses under different lighting conditions, which reduces the stability of local region selection and weakens the reliability of downstream discriminative modeling [8, 24]. Therefore, before mining discriminative local regions, it is desirable to perform a coarse fusion of illumination-consistent local features so as to suppress irrelevant photometric variations while preserving category-related structural cues. Third, most existing attention mechanisms rely on classical similarity computation, such as dot-product attention, to model token interactions [10]. Although effective for general-purpose representation learning, such mechanisms may be insufficient for capturing the complex high-order nonlinear dependencies among subtle local regions in fine-grained recognition, while recent quantum vision studies suggest an alternative route based on parameterized quantum circuits and quantum attention layers [19, 20].

To address the above issues, this paper proposes QGF-Net, a hybrid quantum-classical fine-grained image classification network that integrates self-supervised ViT, an Illumination-Consistent Local Coarse Fusion (ICLCF) module, a Local Discriminative Region Proposal (LDRP) module, a Quantum Measurement Attention (QMA) module, and a Hybrid Quantum-Classical Classifier (HQCC) into a unified framework. Specifically, a self-supervised pretrained ViT is first employed to extract global semantic features and patch token representations from the input image [11]. Then, the ICLCF module is introduced to aggregate semantically related local features under varying lighting conditions, thereby improving the robustness and consistency of local representations. Based on the fused local features, the LDRP module is designed to identify the most informative local regions and narrow the search space for fine-grained recognition. To further enhance subtle local interaction modeling, the QMA module is constructed using parameterized quantum circuits, where qubit measurement outcomes are transformed into adaptive attention weights for reweighting key local regions [18–20]. Finally, the global semantic representation and the quantum-enhanced

local discriminative representation are fused and fed into HQCC to produce the final prediction.

The core idea of QGF-Net is to decompose the FGIC problem into three cooperative stages: stabilizing local representations, selecting discriminative local regions, and modeling complex inter-region dependencies. The ICLCF module is responsible for reducing the sensitivity of local features to lighting variations. The LDRP module further focuses the model on the most category-relevant local patches. On top of this, the QMA module provides an alternative way to model subtle local interactions beyond conventional classical attention. By embedding selected local features into a parameterized quantum circuit and exploiting measurement-derived attention responses, QGF-Net enhances the representation of subtle category-specific cues while maintaining compatibility with classical deep learning backbones [19, 20].

In summary, the main contributions of this paper are as follows. 1. A novel hybrid quantum-classical framework for fine-grained image classification is proposed, which integrates self-supervised ViT, ICLCF, LDRP, QMA, and HQCC into a unified architecture. 2. An ICLCF module is designed to alleviate the instability of local representations caused by illumination variations, shadows, and reflections, thereby improving the robustness of subsequent local discriminative region mining. 3. A QMA module is introduced that uses qubit measurement outcomes to adaptively reweight key local regions, enabling enhanced modeling of subtle and nonlinear dependencies among discriminative local cues. 4. Global semantic features and quantum-enhanced local discriminative features are fused through HQCC, and extensive experiments on multiple fine-grained benchmarks demonstrate the effectiveness of QGF-Net.

2 Related Work

2.1 Fine-Grained Image Classification

Fine-grained image classification (FGIC) has been extensively studied due to its importance in applications such as species recognition, vehicle identification, and medical subtype analysis [2, 3]. Unlike conventional image classification, FGIC focuses on distinguishing subordinate categories with highly similar visual appearances. The major challenge lies in the subtle inter-class differences and large intra-class variations caused by viewpoint, pose, occlusion, background clutter, and illumination changes [8, 24].

Early FGIC methods often relied on strongly supervised annotations, such as bounding boxes, part locations, and keypoints, to localize discriminative object parts and improve classification accuracy. Although these methods achieved promising performance, they required expensive manual annotations and exhibited limited scalability in real-world scenarios [4, 5]. To reduce this dependence on detailed supervision, subsequent studies shifted toward weakly supervised and attention-based methods, where the model was encouraged to automatically discover informative local regions from image-level labels alone [7, 24]. These methods demonstrated that local discriminative cues play a critical role in FGIC and that effective local region modeling is often more important than simply increasing backbone capacity [7, 8].

More recently, researchers have emphasized the importance of jointly modeling global semantics and local discriminative details. Global features provide holistic category context, while local features capture subtle structural or textural differences among visually similar subcategories. As a result, global-local fusion has become a widely adopted paradigm in FGIC [7, 9]. However, despite the progress of existing approaches, accurately identifying stable and discriminative local regions remains challenging, especially when the appearance of local regions is affected by photometric variations such as illumination changes and shadows [8, 24].

2.2 Self-Supervised Vision Transformers for Fine-Grained Recognition

Vision Transformers (ViTs) have recently achieved strong performance across a wide range of visual recognition tasks. By modeling long-range dependencies through self-attention, ViTs provide a flexible mechanism for capturing global contextual information and have shown superior scalability compared with conventional convolutional architectures [10]. At the same time, self-supervised learning has emerged as a powerful paradigm for learning transferable representations from large-scale unlabeled data, thereby reducing the reliance on costly manual annotations [11, 12].

The combination of self-supervised learning and ViT has provided a strong foundation for downstream fine-grained recognition. Self-supervised pretrained ViTs are capable of learning semantically meaningful patch-level representations and robust global features, which are particularly attractive for FGIC where annotated data are often limited [11, 12]. Compared with purely supervised learning, self-supervised initialization generally improves generalization ability and feature transferability [11].

Nevertheless, directly applying self-supervised ViTs to FGIC still has notable limitations. Most self-supervised objectives are designed to encourage global semantic consistency across augmented views, which may bias the model toward holistic object-level representation learning [11]. As a consequence, the learned features may insufficiently emphasize subtle local discriminative regions that are essential for distinguishing fine-grained categories. In addition, while ViT attention maps provide a certain degree of localization ability, they do not explicitly guarantee that the most category-relevant local details will be stably selected under varying imaging conditions [7, 10]. Therefore, enhancing self-supervised ViT with local discriminative region modeling remains an important direction for improving FGIC performance [7, 8].

2.3 Illumination-Robust Local Representation Learning

In practical visual recognition tasks, appearance variations caused by illumination changes, shadows, reflections, and photometric distortions can significantly affect feature stability. This issue is particularly severe in fine-grained classification, where subtle local cues often determine category identity [8, 24]. Even when two local regions are semantically consistent, their feature responses may differ substantially under different lighting conditions, leading to unstable local region selection and degraded discriminative modeling.

To alleviate this problem, prior studies in visual recognition have explored various illumination-robust representation learning strategies, including feature normalization, local contrast enhancement, invariant representation learning, and image or feature fusion mechanisms. These methods aim to suppress photometric variations while preserving structural and semantic information. In fine-grained scenarios, illumination-robust local modeling is especially important because category-relevant differences are often concentrated in small regions with delicate textures and shapes [2, 3].

Although related efforts have improved the robustness of visual representations under varying lighting conditions, most existing methods either focus on general image enhancement or treat local feature stabilization as an auxiliary preprocessing step. They rarely integrate illumination-consistent local fusion with discriminative region proposal in a unified fine-grained classification framework. Moreover, few methods explicitly consider how stabilized local representations can benefit subsequent high-level attention modeling. This motivates us to introduce an illumination-consistent local coarse fusion module before discriminative region mining, so that local feature representations become more robust and reliable for downstream subtle-region analysis.

2.4 Hybrid Quantum-Classical Learning and Quantum Attention

Hybrid quantum-classical learning has recently attracted increasing attention as a promising paradigm for combining the strong representation capability of classical deep neural networks with the expressive potential of parameterized quantum circuits [18, 19, 21]. Instead of replacing classical models entirely, hybrid approaches typically embed compact quantum modules into classical architectures, allowing the model to exploit quantum nonlinear transformations while remaining compatible with current noisy intermediate-scale quantum devices [18, 19].

Existing studies on hybrid quantum-classical learning have explored quantum-enhanced convolutional networks, variational quantum classifiers, and hybrid architectures for image recognition and signal processing tasks. In these approaches, classical networks are often used to extract high-dimensional semantic features, while quantum circuits are employed to further transform compact latent representations [18, 19, 21, 23]. This design is particularly attractive when the quantum module is applied to a reduced set of informative features, since current quantum hardware imposes strict constraints on qubit number and circuit depth [18, 19].

In parallel, recent studies have begun to investigate quantum-inspired or quantum-enhanced attention mechanisms. The key motivation is that conventional attention usually relies on similarity-based formulations, such as dot-product operations, which may be insufficient for modeling subtle and highly nonlinear interactions among informative features. By contrast, parameterized quantum circuits provide an alternative way to encode and transform feature relationships through entanglement and measurement operations [20]. Measurement outcomes from qubits can then be interpreted as adaptive responses for feature reweighting, offering a new perspective on attention modeling [20].

Despite these encouraging developments, the use of hybrid quantum-classical attention for fine-grained image classification remains underexplored. In particular, few existing methods combine self-supervised ViT representations, illumination-robust local feature stabilization, discriminative region selection, and quantum measurement-based attention within a unified framework. Our work aims to fill this gap by designing QGF-Net, in which selected local discriminative regions are refined through a quantum measurement attention mechanism and subsequently fused with global semantic representations for final classification.

3 Method

3.1 Overall Framework

In this paper, we propose QGF-Net, a hybrid quantum-classical framework for fine-grained image classification, which integrates self-supervised Vision Transformer representation learning, illumination-consistent local coarse fusion, discriminative local region mining, quantum measurement-based attention, and adaptive global-local fusion into a unified architecture. The overall goal of QGF-Net is to improve the recognition of fine-grained categories by stabilizing local representations under varying lighting conditions, selecting the most informative local regions, and enhancing their interactions through quantum-inspired feature reweighting, while jointly leveraging global semantic cues and local discriminative details for final prediction [11, 19, 20].

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we first feed it into a self-supervised pretrained Vision Transformer backbone to obtain a global semantic representation and a set of patch-level local token representations. Specifically, the image is divided into N non-overlapping patches, which are linearly projected into token embeddings and processed by the ViT encoder [10]. Let the output token sequence be denoted as

$$T = \{t_{cls}, t_1, t_2, \dots, t_N\}, \quad (1)$$

where t_{cls} is the class token representing global semantic information, and $t_i \in \mathbb{R}^d$ denotes the feature representation of the i -th image patch.

Although self-supervised ViT can provide strong semantic representations, the patch tokens may still be sensitive to local appearance variations caused by illumination changes, shadows, and reflections. Such photometric disturbances can reduce the consistency of semantically similar local regions and negatively affect the subsequent selection of discriminative patches [8, 24]. To address this issue, we introduce an Illumination-Consistent Local Coarse Fusion (ICLCF) module, which performs local feature aggregation on patch tokens to enhance their robustness under varying lighting conditions. The fused local token set is denoted as

$$\tilde{T} = \{\tilde{t}_1, \tilde{t}_2, \dots, \tilde{t}_N\}. \quad (2)$$

Based on the fused token representations, we further construct a Local Discriminative Region Proposal (LDRP) module to estimate the importance of each local region and select the top- K most informative patch tokens for fine-grained recognition

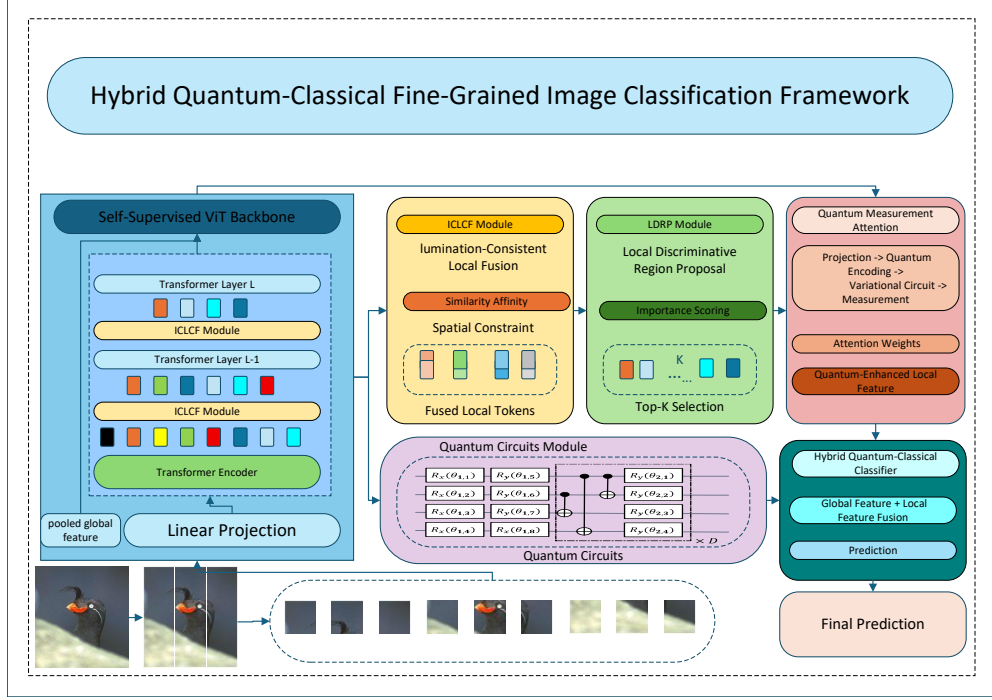


Fig. 1 Overview of QGF-Net. Built upon a ViT backbone, QGF-Net consists of a fusion branch and a local discriminative branch. The fusion branch employs the ICLCF module to improve the robustness and consistency of local token features under varying illumination, while the local discriminative branch uses the LDRP module to identify the most informative local regions. The QMA module then models the dependencies among key local regions through lightweight parameterized quantum circuits, and the resulting features are finally fed into HQCC for prediction

[4, 5, 7]. These selected local features are expected to contain subtle category-specific cues, such as texture, edge shape, and structural patterns. Let the selected local feature subset be denoted as

$$T_K = \{\tilde{t}_{i_1}, \tilde{t}_{i_2}, \dots, \tilde{t}_{i_K}\}. \quad (3)$$

To further model the complex dependencies among the selected local discriminative regions, we design a Quantum Measurement Attention (QMA) module. Instead of relying solely on classical dot-product attention, QMA embeds the selected local features into a parameterized quantum circuit and uses qubit measurement outcomes to generate adaptive attention scores [18–20]. In this way, the model can capture subtle nonlinear interactions among local discriminative regions. The resulting quantum-enhanced local representation is denoted as

$$f_{local}^q. \quad (4)$$

Meanwhile, the global semantic representation extracted by the self-supervised ViT backbone is preserved as

$$f_{global} = t_{cls}. \quad (5)$$

Finally, the global representation and the quantum-enhanced local representation are fused and fed into a Hybrid Quantum-Classical Classifier (HQCC) for final category prediction. The fusion process can be expressed as

$$f_{fusion} = \phi([f_{global}; f_{local}^q]), \quad (6)$$

where $[\cdot; \cdot]$ denotes feature concatenation and $\phi(\cdot)$ denotes a learnable projection function. The fused feature is then passed through the classifier to produce the probability distribution over fine-grained categories:

$$\hat{y} = \text{Softmax}(g(f_{fusion})), \quad (7)$$

where $g(\cdot)$ denotes the classification head and $\hat{y}$ is the predicted label distribution.

QGF-Net can therefore be viewed as a progressive fine-grained recognition pipeline composed of four consecutive stages: (1) global and local feature extraction through self-supervised ViT, (2) illumination-consistent local feature stabilization through ICLCF, (3) discriminative local region selection through LDRP, and (4) higher-order local interaction modeling through QMA and final global-local fusion classification. Compared with existing fine-grained classification methods, QGF-Net has three main characteristics. First, it introduces an illumination-aware local fusion mechanism to improve the robustness of local feature representations before local region mining. Second, it explicitly separates the tasks of feature stabilization and discriminative region selection, thereby improving the reliability of local patch discovery. Third, it incorporates a quantum measurement-based attention module to enhance the modeling of subtle local dependencies beyond conventional classical attention mechanisms [7, 20]. These design choices make QGF-Net particularly suitable for fine-grained image classification, where stable local details and subtle inter-region interactions are both crucial.

3.1.1 Motivation

In fine-grained image classification, category-specific cues often reside in subtle local regions, such as feather texture, beak shape, wing pattern, or tail contour [4, 5, 8]. However, these local discriminative cues are highly sensitive to illumination changes, shadows, reflections, and background interference. Even when two local patches correspond to semantically consistent object parts, their feature responses may differ significantly under different imaging conditions. As a result, directly selecting discriminative local regions from raw patch tokens may lead to unstable region proposals and degraded recognition performance [8, 24].

To alleviate this issue, we propose an Illumination-Consistent Local Coarse Fusion (ICLCF) module, which is inserted between the self-supervised ViT backbone and the local discriminative region proposal stage. The objective of ICLCF is not to reconstruct or enhance the original image, but to perform a coarse aggregation of

semantically related local token features so that photometric disturbances can be suppressed while category-relevant structural information can be preserved. In this way, the subsequent discriminative region proposal module can operate on more stable local representations.

3.1.2 Formulation

Let the patch token sequence extracted by the ViT backbone be

$$T_p = \{t_1, t_2, \dots, t_N\}, \quad t_i \in \mathbb{R}^d. \quad (8)$$

For each local token t_i , we construct a local neighborhood $\mathcal{N}(i)$ according to feature similarity and spatial proximity. The motivation is that semantically related local patches under different lighting conditions should still exhibit partial consistency in feature space and image structure. Therefore, instead of treating each token independently, we aggregate the neighboring tokens to obtain a more robust local representation.

We first compute an affinity score between the i -th token and its neighboring token t_j as

$$s_{ij} = \frac{W_q t_i \cdot W_k t_j}{\sqrt{d}} + \lambda \rho(i, j), \quad (9)$$

where W_q and W_k are learnable projection matrices, and $\rho(i, j)$ denotes a spatial consistency term that measures the relative proximity between patch i and patch j . The coefficient λ controls the contribution of spatial locality. This formulation allows the module to jointly consider feature similarity and structural continuity during local aggregation.

The normalized aggregation weight is then obtained by

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{j' \in \mathcal{N}(i)} \exp(s_{ij'})}. \quad (10)$$

Based on these weights, the illumination-consistent fused token $\tilde{t}_i$ is computed as

$$\tilde{t}_i = t_i + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W_v t_j, \quad (11)$$

where W_v is a learnable value projection matrix. The residual connection preserves the original local semantics, while the weighted aggregation incorporates neighboring contextual cues that are consistent in structure but potentially different in photometric appearance.

After local aggregation, we further introduce a lightweight feature normalization and gating mechanism to suppress illumination-sensitive responses. The normalized fused token is expressed as

$$\hat{t}_i = \text{LayerNorm}(\tilde{t}_i), \quad (12)$$

and the final output of the ICLCF module is

$$t_i^f = \gamma_i \odot \hat{t}_i + (1 - \gamma_i) \odot t_i, \quad (13)$$

where $\odot$ denotes element-wise multiplication and $\gamma_i \in [0, 1]^d$ is an adaptive gating vector generated by a lightweight multilayer perceptron:

$$\gamma_i = \sigma(\text{MLP}([t_i; \hat{t}_i])), \quad (14)$$

with $\sigma(\cdot)$ denoting the sigmoid activation function. This gating mechanism allows the model to adaptively determine how much fused contextual information should be injected into each local token.

The final output of the ICLCF module is therefore

$$T_f = \{t_1^f, t_2^f, \dots, t_N^f\}. \quad (15)$$

These fused tokens are subsequently fed into the local discriminative region proposal module for selecting the most informative local patches.

3.1.3 Interpretation

The proposed ICLCF module can be interpreted as a local feature stabilization mechanism tailored for fine-grained recognition. Different from standard self-attention, which is mainly designed for generic token interaction modeling, ICLCF focuses specifically on reducing the inconsistency of semantically similar local tokens under illumination variation. It performs coarse fusion rather than full feature replacement, which preserves the original discriminative cues while enhancing their robustness.

From the perspective of fine-grained recognition, ICLCF provides two main benefits. First, it improves the consistency of local token features by aggregating neighboring cues with similar semantics and structural context. This reduces the risk that important local parts are overlooked due to local photometric noise. Second, it prepares more stable token representations for subsequent discriminative region proposal, thereby increasing the reliability of selecting category-relevant local patches [8, 24].

From the perspective of the overall framework, ICLCF serves as a bridge between generic self-supervised representation learning and task-specific subtle-region discrimination. The self-supervised ViT backbone provides rich semantic patch representations, while ICLCF refines these representations into illumination-consistent local features that are more suitable for downstream fine-grained region mining and quantum attention modeling.

3.2 Local Discriminative Region Proposal Module

3.2.1 Motivation

Although the ICLCF module improves the robustness of local token representations under varying illumination conditions, not all local patches contribute equally to fine-grained recognition [4, 5, 7]. In fine-grained image classification, only a small subset of

local regions usually contains category-discriminative cues, such as subtle texture differences, local shape variations, and fine structural patterns. Many other local tokens correspond to background, redundant object parts, or weakly informative regions. If all local patches are treated equally, the discriminative information may be diluted, which limits the ability of the subsequent attention and classification modules to focus on truly informative details.

To address this issue, we introduce a Local Discriminative Region Proposal (LDRP) module, whose objective is to identify and preserve the most informative local regions for downstream subtle-region modeling. Instead of directly applying attention to all fused local tokens, LDRP estimates the discriminative importance of each token and selects the top- K local regions with the highest relevance to fine-grained category recognition. In this way, the search space of local modeling is effectively narrowed, and the subsequent quantum measurement attention module can concentrate on a compact set of highly informative region features.

3.2.2 Discriminative Region Scoring

Let the output of the ICLCF module be denoted as

$$T_f = \{t_1^f, t_2^f, \dots, t_N^f\}, \quad t_i^f \in \mathbb{R}^d. \quad (16)$$

For each fused local token t_i^f , we estimate a discriminative score that reflects its importance for fine-grained classification. Since local discriminative cues should be evaluated not only from local appearance itself but also in relation to the global semantic context, we incorporate the global class token t_{cls} into the scoring process. Specifically, we first construct a context-aware token representation by concatenating the local token with the global semantic feature:

$$z_i = [t_i^f; t_{cls}], \quad (17)$$

where $[\cdot; \cdot]$ denotes feature concatenation. The discriminative score for the i -th local token is then computed by a lightweight multilayer perceptron:

$$r_i = W_2 \delta(W_1 z_i + b_1) + b_2, \quad (18)$$

where W_1 and W_2 are learnable projection matrices, b_1 and b_2 are bias terms, and $\delta(\cdot)$ denotes a nonlinear activation function such as GELU or ReLU. To normalize the region importance across all local tokens, we apply a softmax operation:

$$\beta_i = \frac{\exp(r_i)}{\sum_{j=1}^N \exp(r_j)}. \quad (19)$$

Here, β_i can be interpreted as the relative discriminative importance of the i -th patch token. Higher values indicate that the corresponding local region is more likely to contain subtle category-specific cues.

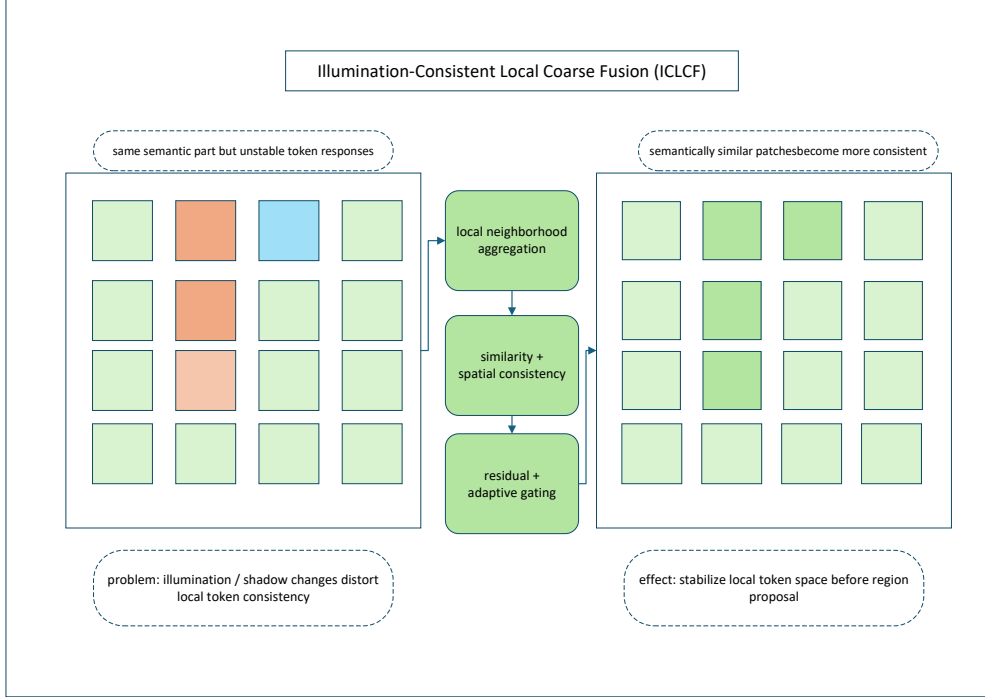


Fig. 2 Overview of the Illumination-Consistent Local Coarse Fusion (ICLCF) module. It mitigates inconsistent local token responses caused by illumination variations by aggregating neighborhood features, enforcing similarity and spatial consistency, and applying adaptive gating, resulting in more stable and consistent local representations

3.2.3 Top- K Region Selection

After obtaining the token importance scores $\{\beta_i\}_{i=1}^N$, we rank all local tokens and select the top- K tokens with the highest discriminative scores. The selected token index set is defined as

$$\mathcal{I}_K = \text{TopK}(\{\beta_i\}_{i=1}^N, K), \quad (20)$$

and the corresponding local discriminative token subset is

$$T_K = \{t_{i_1}^f, t_{i_2}^f, \dots, t_{i_K}^f\}, \quad i_m \in \mathcal{I}_K. \quad (21)$$

To preserve both semantic information and selection confidence, we further weight the selected local features by their normalized importance scores:

$$\hat{t}_{i_m} = \beta_{i_m} \cdot t_{i_m}^f, \quad m = 1, 2, \dots, K. \quad (22)$$

Thus, the output of the LDRP module can be represented as

$$\hat{T}_K = \{\hat{t}_{i_1}, \hat{t}_{i_2}, \dots, \hat{t}_{i_K}\}. \quad (23)$$

This weighted top- K selection strategy enables QGF-Net to preserve the strongest local cues while suppressing less informative or potentially noisy regions.

3.2.4 Spatial Awareness and Diversity Constraint

A potential limitation of naive top- K region selection is that multiple selected tokens may come from highly overlapping or spatially redundant regions. In fine-grained recognition, such redundancy can reduce the diversity of local evidence and weaken the coverage of multiple discriminative object parts [4, 5]. To alleviate this issue, we further introduce a lightweight diversity-aware constraint during region proposal.

Let $p_i = (x_i, y_i)$ denote the spatial position of the i -th patch token in the patch grid. For two selected local regions i and j , we define their spatial redundancy as

$$d_{ij} = \exp\left(-\frac{\|p_i - p_j\|_2^2}{\tau}\right), \quad (24)$$

where τ is a scale factor controlling spatial sensitivity. During token selection, we penalize highly redundant local regions by modifying the region score as

$$\tilde{\beta}_i = \beta_i - \mu \sum_{j \in S} d_{ij}, \quad (25)$$

where S denotes the set of already selected tokens and μ is a balancing coefficient. In practice, this mechanism encourages the selected top- K local regions to cover multiple informative object parts rather than collapsing to a single highly activated area.

3.3 Quantum Measurement Attention Module

3.3.1 Motivation

After the LDRP module selects the most informative local regions, a remaining challenge is how to effectively model the subtle relationships among these local discriminative cues. In fine-grained image classification, category differences are often determined not by a single local part, but by the joint configuration of multiple subtle regions [4, 5, 7]. Therefore, it is essential to model inter-region dependencies rather than treating selected local regions independently.

Conventional attention mechanisms typically rely on dot-product similarity or linear projections to estimate token interactions [10]. While effective in many visual recognition tasks, these operations may be limited in capturing complex high-order nonlinear dependencies among subtle local regions. To overcome this limitation, we introduce a Quantum Measurement Attention (QMA) module, which employs a parameterized quantum circuit to model the interactions among the selected local discriminative regions [18–20]. By encoding compact local features into quantum states and transforming them through learnable quantum operations, the proposed module generates measurement-based attention responses that are subsequently used to reweight local region features.

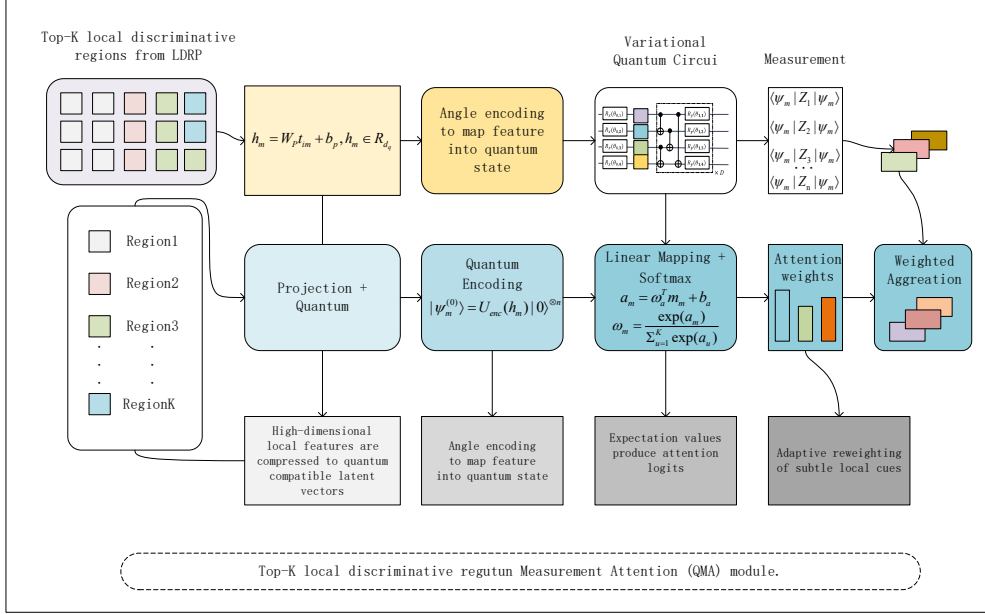


Fig. 3 Overview of the proposed Quantum Measurement Attention (QMA) module. Given the top- K local discriminative features selected by LDRP, each local feature is first projected into a compact latent space and normalized for quantum-compatible encoding. The resulting feature vectors are then mapped into quantum states through angle encoding and processed by a learnable variational quantum circuit. After measurement, the expectation values of Pauli-Z observables are converted into attention logits and normalized by a softmax function to produce attention weights. Finally, the selected local features are adaptively aggregated according to the quantum measurement-based attention, yielding a quantum-enhanced local representation

3.3.2 Local Feature Preparation

Let the selected top- K local discriminative features from LDRP be

$$\hat{T}_K = \{\hat{t}_{i_1}, \hat{t}_{i_2}, \dots, \hat{t}_{i_K}\}, \quad \hat{t}_{i_m} \in \mathbb{R}^d. \quad (26)$$

Since the dimensionality of transformer features is usually much higher than the number of qubits available in practical quantum circuits, we first project each selected local feature into a compact latent representation:

$$h_m = W_p \hat{t}_{i_m} + b_p, \quad h_m \in \mathbb{R}^{d_q}, \quad (27)$$

where $W_p \in \mathbb{R}^{d_q \times d}$ is a learnable projection matrix, b_p is a bias vector, and d_q is the reduced feature dimension compatible with quantum encoding. Typically, d_q is chosen to match the number of qubit rotation parameters used in the quantum circuit [18, 19]. To facilitate stable encoding, we further normalize the projected feature vector:

$$\bar{h}_m = \text{Norm}(h_m). \quad (28)$$

3.3.3 Quantum Encoding and Variational Circuit

For each selected local region feature $\bar{h}_m$, we construct a quantum state through angle encoding. Assuming an n -qubit quantum circuit, the encoded state is initialized as

$$|\psi_m^{(0)}\rangle = U_{enc}(\bar{h}_m)|0\rangle^{\otimes n}, \quad (29)$$

where $U_{enc}(\cdot)$ denotes the quantum feature encoding operation and $|0\rangle^{\otimes n}$ is the initial all-zero quantum state [18, 19].

After quantum encoding, the state is passed through a learnable variational quantum circuit:

$$|\psi_m\rangle = U_\theta|\psi_m^{(0)}\rangle, \quad (30)$$

where U_θ is composed of multiple layers of trainable single-qubit rotations and entangling gates such as CNOT or CZ gates. The role of this parameterized quantum circuit is to model nonlinear dependencies in the latent feature space and produce expressive transformations of the local discriminative representations [18–20].

3.3.4 Measurement-Based Attention Generation

Once the encoded local feature passes through the variational quantum circuit, the transformed quantum state is measured to obtain a classical response vector. Specifically, the expectation value of the Pauli-Z observable on each qubit is computed as

$$m_m = [\langle\psi_m|Z_1|\psi_m\rangle, \langle\psi_m|Z_2|\psi_m\rangle, \dots, \langle\psi_m|Z_n|\psi_m\rangle]. \quad (31)$$

We then map the measurement vector to a scalar attention logit:

$$a_m = w_a^\top m_m + b_a, \quad (32)$$

where w_a and b_a are learnable parameters. The normalized quantum attention weight for the m -th selected local region is obtained by

$$\omega_m = \frac{\exp(a_m)}{\sum_{u=1}^K \exp(a_u)}. \quad (33)$$

The final quantum-enhanced local feature is computed as a weighted aggregation of the selected local features:

$$f_{local}^q = \sum_{m=1}^K \omega_m \hat{t}_{i_m}. \quad (34)$$

3.4 Hybrid Quantum-Classical Classifier

3.4.1 Motivation

After obtaining the global semantic representation from the self-supervised ViT backbone and the quantum-enhanced local discriminative representation from the QMA module, the final step is to effectively combine these complementary sources of information for fine-grained classification [7, 9]. In FGIC, global semantic information

provides holistic object-level context, while local discriminative information captures subtle region-specific cues that are crucial for distinguishing visually similar subcategories. Relying only on global features may overlook fine-grained local differences, whereas using only local cues may weaken semantic consistency and global contextual reasoning. Therefore, a joint global-local classification strategy is essential [7, 8].

3.4.2 Global-Local Feature Fusion

Let the global semantic representation obtained from the ViT class token be denoted as

$$f_{global} \in \mathbb{R}^d, \quad (35)$$

and the quantum-enhanced local representation generated by QMA be denoted as

$$f_{local}^q \in \mathbb{R}^d. \quad (36)$$

To fully exploit the complementary properties of these two representations, we first concatenate them into a joint feature vector:

$$f_{concat} = [f_{global}; f_{local}^q] \in \mathbb{R}^{2d}. \quad (37)$$

A learnable fusion projection is then introduced:

$$f_{fusion} = \phi(f_{concat}) = \delta(W_f f_{concat} + b_f), \quad (38)$$

where W_f and b_f are learnable parameters and $\delta(\cdot)$ is a nonlinear activation function such as GELU.

To further enhance the flexibility of feature fusion, we incorporate an adaptive global-local reweighting mechanism:

$$[\lambda_g, \lambda_l] = \text{Softmax}(W_r[f_{global}; f_{local}^q] + b_r), \quad (39)$$

where $\lambda_g + \lambda_l = 1$. Based on these coefficients, the adaptive fusion feature can be expressed as

$$f_{adaptive} = \lambda_g f_{global} + \lambda_l f_{local}^q. \quad (40)$$

The final fusion feature is written as

$$f_{final} = f_{fusion} + W_a f_{adaptive}. \quad (41)$$

The final fused feature f_{final} is fed into a classical classification head to predict the category label:

$$\hat{y} = \text{Softmax}(W_c f_{final} + b_c). \quad (42)$$

3.5 Hybrid Quantum-Classical Classifier

To effectively train QGF-Net, we adopt a multi-objective loss function:

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{local} + \lambda_3 L_{align} + \lambda_4 L_{qreg}, \quad (43)$$

where L_{cls} is the classification loss, L_{local} is the local discriminative learning loss, L_{align} is the global-local alignment loss, and L_{qreg} is the quantum attention regularization loss.

The standard cross-entropy loss is used as

$$L_{cls} = - \sum_{c=1}^C y_c \log \hat{y}_c. \quad (44)$$

A local center loss is used to improve the separability of local representations:

$$L_{local} = \|f_{local}^q - c_y\|_2^2. \quad (45)$$

To ensure semantic consistency between the global and local branches, a cosine alignment term is introduced:

$$L_{align} = 1 - \frac{f_{global}^\top f_{local}^q}{\|f_{global}\|_2 \|f_{local}^q\|_2}. \quad (46)$$

For QMA, an entropy-based regularization is adopted:

$$L_{qreg} = - \sum_{m=1}^K \omega_m \log \omega_m. \quad (47)$$

The overall optimization objective is therefore

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{local} + \lambda_3 L_{align} + \lambda_4 L_{qreg}. \quad (48)$$

4 Experiments and Results Analysis

4.1 Experimental Setup

4.1.1 Datasets and Evaluation Metrics

To validate the effectiveness and generalization of QGF-Net, experiments are conducted on three representative fine-grained benchmarks: CUB-200-2011, Stanford Cars, and FGVC-Aircraft [13, 15, 16].

CUB-200-2011 contains 200 bird subcategories and 11,788 images, together with multiple forms of annotations, including class labels, bounding boxes, part annotations, and attributes [13]. Following the standard split, 5,994 images are used for training and 5,794 images are used for testing. Stanford Cars contains 196 car categories and 16,185 images, including 8,144 training images and 8,041 testing images [15]. FGVC-Aircraft contains 100 aircraft variants with 10,000 images following the standard train/validation/test split, and the standard training split is used for model fitting while the test split is used for final evaluation [16].

In this work, only image-level class labels are used for training and evaluation, without introducing additional supervision such as bounding boxes or keypoint annotations. Since these datasets exhibit subtle inter-class differences as well as significant intra-class variations in pose, illumination, and background, they are particularly suitable for evaluating methods that focus on discriminative local region modeling and global-local feature fusion [2, 3].

It should also be noted that the official description of CUB-200-2011 indicates a potential overlap between some images in this dataset and the ImageNet/Flickr pretraining corpora [14]. Therefore, when interpreting results obtained with ImageNet-supervised or large-scale web-pretrained models, caution should be taken to avoid overstating the contribution of QGF-Net alone.

For evaluation, Top-1 Accuracy is adopted as the primary metric to measure the overall classification performance on the test set. In addition, Macro-F1 is reported as an auxiliary metric to assess the balanced classification capability across categories. To further evaluate the practicality of the model, the number of parameters (Params) and floating-point operations (FLOPs) are also reported to reflect model complexity. For robustness analysis, classification accuracy under different perturbation settings and the average performance drop are further measured to evaluate model stability under challenging conditions such as illumination variation, shadow interference, and partial occlusion.

4.1.2 Implementation Details

Following the design presented in Section 3, Vision Transformer (ViT) is adopted as the backbone network, and the model is initialized with DINOv2 self-supervised pretrained weights [10, 11]. All input images are resized to 224×224 . During training, standard data augmentation strategies are employed, including random cropping, horizontal flipping, and normalization. The optimizer is AdamW [17, 24].

The training process is divided into two stages. In the first stage, part of the early backbone layers is frozen, and only the newly introduced modules, including ICLCF, LDRP, QMA, and HQCC, are trained so as to preserve the stable representations learned during self-supervised pretraining. In the second stage, the whole network is fine-tuned end-to-end to further improve the compatibility between the backbone features and the local discriminative enhancement modules. This two-stage strategy helps improve optimization stability, especially in QGF-Net with quantum measurement attention [11, 19].

For module configuration, the number of candidate local discriminative regions in LDRP is selected from $K \in \{4, 8, 16\}$ and further analyzed in the ablation study. For QMA, a lightweight parameterized quantum circuit is adopted. The default setting uses 4 qubits, 2 variational layers, and a reduced feature dimension of $d_q = 16$ [18, 19]. Inference latency and FPS are measured under a fixed 224×224 inference setting with batch size 1.

4.2 Comparison Methods

To systematically verify the effectiveness of QGF-Net, representative baselines from different categories are selected for comparison. The first category consists of conventional convolutional baselines, represented by ResNet-50, which reflect the performance of models without explicit local discriminative region modeling [1]. The second category includes Transformer-based backbone baselines, such as standard ViT-B/16 and a self-supervised ViT baseline, which are used to evaluate the performance of strong global representation backbones in fine-grained scenarios [10, 11]. The third category contains locally enhanced baselines, including a classical attention variant and a variant equipped with ICLCF and LDRP, in order to examine the contribution of explicit local discriminative modeling. The fourth category is the full model proposed in this work, which integrates ICLCF, LDRP, QMA, and HQCC into QGF-Net.

4.3 Main Experimental Results and Analysis

4.3.1 Quantitative Comparison Results

Table 1 Comparison of QGF-Net with representative baselines.

Method	Pretraining	Local Mod.	Top-1 (%)	Macro-F1 (%)	Params (M)	FLOPs (G)
A ¹	ImageNet supervised	No	86.3	84.8	25.6	4.1
B ²	ImageNet supervised	No	88.0	86.7	86.6	16.9
C ³	DINOv2	No	89.4	88.2	86.0	21.0
D ⁴	DINOv2	Yes	90.1	89.0	87.2	21.7
E ⁵	DINOv2	Yes	90.9	89.9	88.1	22.4
F ⁶	DINOv2	Yes	92.0	91.2	89.3	23.6

¹A = ResNet-50.

²B = ViT-B/16.

³C = Self-Supervised ViT Baseline.

⁴D = Self-Supervised ViT + Classical Attention.

⁵E = Self-Supervised ViT + ICLCF + LDRP.

⁶F = QGF-Net.

As shown in Table 1, QGF-Net achieves the best Top-1 Accuracy and Macro-F1 among all compared models. Compared with ResNet-50, QGF-Net improves Top-1 Accuracy by 5.7 percentage points and Macro-F1 by 6.4 percentage points, indicating that the Transformer-based global-local collaborative framework is significantly more effective than conventional convolutional global classification for fine-grained recognition. Compared with the standard ViT-B/16 backbone, QGF-Net improves Top-1 Accuracy and Macro-F1 by 4.0 and 4.5 percentage points, respectively. This suggests that even with a Transformer backbone, the absence of explicit fine-grained local discriminative modeling limits the ability of the model to capture subtle category-specific cues.

Furthermore, compared with the self-supervised ViT baseline, QGF-Net improves Top-1 Accuracy by 2.6 percentage points and Macro-F1 by 3.0 percentage points. This result demonstrates that even though DINOv2 provides strong self-supervised global semantic representations, fine-grained image classification still heavily depends on explicit modeling of discriminative local regions. Compared with the classical local attention variant, QGF-Net further improves Top-1 Accuracy and Macro-F1 by 1.9 and 2.2 percentage points, respectively. In comparison with the variant using only ICLCF and LDRP, the full model still obtains gains of 1.1 percentage points in Top-1 Accuracy and 1.3 percentage points in Macro-F1, indicating that the introduction of QMA and HQCC does not merely increase model complexity, but indeed enhances the representation of high-value local regions and the final decision-making capability.

4.3.2 Multi-Dataset Comparison

The results across multiple datasets are summarized in Table 2.

Table 2 Comparison across CUB-200-2011, Stanford Cars, and FGVC-Aircraft.

Method	CUB (%)	Cars (%)	Aircraft (%)	Avg. (%)
ResNet-50	86.3	91.1	84.6	87.3
ViT-B/16	88.0	92.0	86.6	88.9
Self-Supervised ViT Baseline	89.4	92.3	86.9	89.5
+ Classical Attention	90.1	93.0	87.5	90.2
+ ICLCF + LDRP	90.9	93.6	88.2	90.9
QGF-Net	92.0	94.2	89.5	91.9

QGF-Net achieves the best results on all three datasets. The gains are consistent across categories with different structures and visual patterns, indicating that QGF-Net is not limited to a single object type. In particular, the performance gain on FGVC-Aircraft suggests that the proposed local stabilization and local interaction modeling strategy is especially beneficial in scenarios where category differences are subtle, spatially sparse, and structurally entangled.

4.3.3 Discussion with Recent Fine-Grained Classification Results

Recent public studies indicate that Transformer-based fine-grained classification models with local enhancement mechanisms have generally reached strong performance levels on representative benchmarks [2, 7, 9]. Against this background, QGF-Net achieves competitive results while maintaining a clearly interpretable processing pipeline consisting of local stabilization, discriminative region selection, quantum measurement enhancement, and global-local fusion. More importantly, the contribution of this work does not lie in simply adopting a larger or stronger backbone. Instead, QGF-Net demonstrates that even with a strong self-supervised backbone, fine-grained classification performance can still be further improved through structured local modeling rather than merely scaling up pretraining resources.

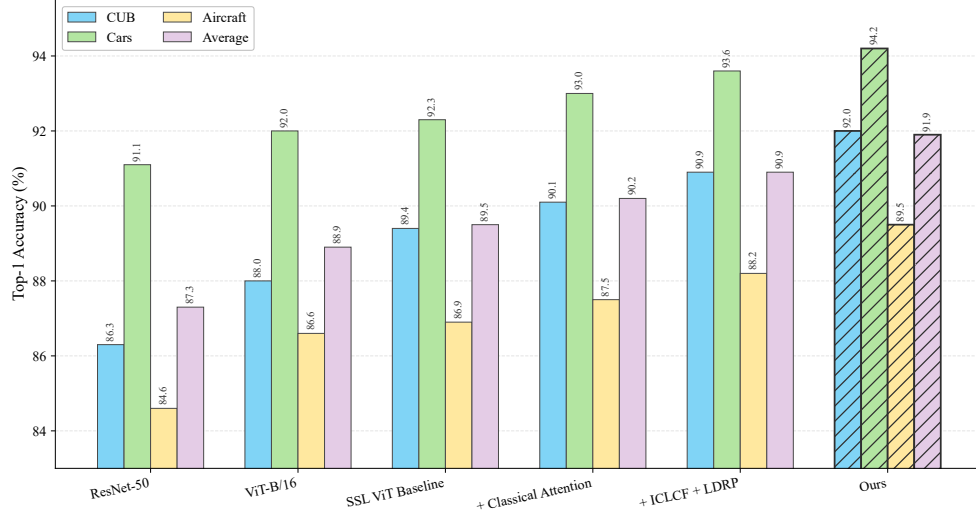


Fig. 4 Comparison of Top-1 accuracy (%) on three datasets (CUB, Cars, and Aircraft). The compared methods range from ResNet-50, ViT-B/16, and the self-supervised ViT baseline to models equipped with Classical Attention, ICLCF+LDRP, and QGF-Net. Among all methods, QGF-Net achieves the best performance on each dataset and also obtains the highest average Top-1 accuracy

4.4 Robustness Experiments and Analysis

To verify the stability of QGF-Net under complex illumination and partial occlusion, robustness evaluations are conducted under brightness perturbation, shadow perturbation, and occlusion perturbation. The results are shown in Table 3.

Table 3 Robustness comparison under different perturbation settings.

Method	Clean	Brightness	Shadow	Occlusion	Avg. Drop
Self-Supervised ViT Baseline	89.4	82.4	80.8	78.1	9.0
+ Classical Attention	90.1	84.0	82.5	80.2	7.9
QGF-Net	92.0	87.0	85.6	83.4	6.0

As shown in Table 3, QGF-Net achieves the best performance under all perturbation conditions. Under brightness variation, QGF-Net reaches 87.0%, which is 4.6 percentage points higher than the self-supervised ViT baseline. Under shadow perturbation, it achieves 85.6%, improving by 4.8 percentage points, and under partial occlusion, it obtains 83.4%, improving by 5.3 percentage points. More importantly, the average performance drop of QGF-Net is only 6.0%, which is substantially lower than 9.0% for the self-supervised ViT baseline and 7.9% for the classical attention variant. This demonstrates that QGF-Net not only delivers higher accuracy under standard

testing conditions, but also maintains stronger robustness under challenging imaging scenarios.

For completeness, the relative drops are also summarized in Table 4.

Table 4 Relative accuracy drop under perturbations.

Method	Brightness Drop (%)	Shadow Drop (%)	Occlusion Drop (%)	Avg. Drop (%)
A ¹	7.0	8.6	11.3	9.0
B ²	6.1	7.6	9.9	7.9
C ³	5.0	6.4	8.6	6.0

¹A = Self-Supervised ViT Baseline.

²B = + Classical Attention.

³C = QGF-Net.

4.5 Ablation Study

4.5.1 Effectiveness of Each Module

To further analyze the contribution of each component to the final performance, systematic ablation experiments are conducted, as shown in Table 5.

Table 5 Ablation results of different modules.

Model Variant	ICLCF	LDRP	QMA	HQCC	Top-1 Acc. (%)	Macro-F1 (%)
A ¹	×	×	×	×	89.4	88.2
Variant-1	✓	×	×	×	90.0	88.9
Variant-2	✓	✓	×	×	90.8	89.7
Variant-3	✓	✓	Classical Attention	×	91.1	90.0
Variant-4	✓	✓	QMA	×	91.6	90.7
Full Model	✓	✓	QMA	✓	92.0	91.2

¹A = Baseline (Self-Supervised ViT).

It can be observed from Table 5 that all modules contribute positively to the final performance. First, after adding ICLCF to the self-supervised ViT baseline, Top-1 Accuracy increases from 89.4% to 90.0%, and Macro-F1 rises from 88.2% to 88.9%. This shows that in fine-grained recognition, stabilizing local representations before selecting discriminative regions is more reliable than directly extracting local regions from raw patch tokens.

After further introducing LDRP, performance increases to 90.8% and 89.7%, which are 0.8 percentage points higher than Variant-1 on both metrics. This indicates that even after ICLCF stabilization, not all local regions carry equal discriminative value.

By explicitly selecting the most informative local regions, the model can focus more effectively on truly critical fine-grained cues.

When QMA is replaced by classical attention, Variant-3 achieves 91.1% Top-1 Accuracy and 90.0% Macro-F1, whereas Variant-4 equipped with QMA reaches 91.6% and 90.7%, improving by 0.5 and 0.7 percentage points, respectively. After further adding HQCC for adaptive global-local fusion, the full model reaches 92.0% Top-1 Accuracy and 91.2% Macro-F1, which are 0.4 and 0.5 percentage points higher than Variant-4.

Table 6 Incremental contribution of each module.

Variant	Top-1 Acc. (%)	Gain over Previous	Macro-F1 (%)	Gain over Previous
Baseline	89.4	—	88.2	—
+ ICLCF	90.0	+0.6	88.9	+0.7
+ LDRP	90.8	+0.8	89.7	+0.8
+ Classical Attention	91.1	+0.3	90.0	+0.3
+ QMA	91.6	+0.5	90.7	+0.7
+ HQCC	92.0	+0.4	91.2	+0.5

4.5.2 Effect of the Number of Local Regions K

To study the impact of the number of local regions selected by LDRP, experiments are conducted with $K = 4, 8, 16$. The results are reported in Table 7.

Table 7 Performance comparison under different values of K .

K	4	8	16
Top-1 Accuracy (%)	91.2	92.0	91.7
Macro-F1 (%)	90.3	91.2	90.9

As shown in Table 7, the best performance is achieved when $K = 8$. When $K = 4$, the model retains only a very small number of local regions and may miss some auxiliary discriminative cues. When $K = 16$, more local regions are preserved, but this also introduces additional redundant and weak-information regions. These results indicate that fine-grained classification does not benefit from selecting as many local regions as possible. Instead, a balance must be achieved between preserving sufficient discriminative information and suppressing local redundancy.

4.5.3 Sensitivity Analysis of the Quantum Circuit Configuration

To analyze the sensitivity of QMA to quantum module configuration, additional experiments are conducted with different numbers of qubits, circuit layers, and reduced feature dimensions. The results are shown in Table 8.

Table 8 Sensitivity analysis of QMA configurations.

Qubits n	Layers L	Reduced Dim d_q	Top-1 Accuracy (%)	Macro-F1 (%)	Remarks
2	1	8	90.8	89.8	(1) ¹
4	1	16	91.4	90.4	(2) ²
4	2	16	92.0	91.2	(3) ³
6	2	32	92.2	91.3	(4) ⁴

¹(1) Limited capacity.

²(2) Basic configuration.

³(3) Recommended configuration.

⁴(4) Slightly better but less stable.

As shown in Table 8, model performance generally improves as the capacity of the quantum circuit increases, but the gain is not linear. When the configuration is increased from 2 qubits and 1 layer to 4 qubits and 2 layers, Top-1 Accuracy rises from 90.8% to 92.0%. However, when the circuit is further enlarged to 6 qubits, 2 layers, and $d_q = 32$, the highest accuracy reaches 92.2%, but training becomes less stable and the reproducibility cost increases. Therefore, considering accuracy, stability, and implementation complexity together, the configuration of 4 qubits, 2 layers, and $d_q = 16$ is selected as the recommended setting.

4.6 Classical Attention versus QMA

To isolate the contribution of QMA, further comparison is conducted under the same local modeling pipeline, as shown in Table 9.

Table 9 Comparison between classical attention and QMA.

Local Modeling Strategy	Top-1 Accuracy (%)	Macro-F1 (%)	Params (M)	FLOPs (G)
Strategy A ¹	91.1	90.0	88.5	22.8
Strategy B ²	91.6	90.7	88.9	23.2
Strategy C ³	92.0	91.2	89.3	23.6

¹Strategy A = ICLCF + LDRP + Classical Attention.

²Strategy B = ICLCF + LDRP + QMA.

³Strategy C = ICLCF + LDRP + QMA + HQCC.

Compared with classical attention, QMA brings consistent gains with only a modest increase in parameters and FLOPs. This suggests that the performance improvement is not primarily caused by a large increase in model scale, but by better modeling of subtle local region relationships. HQCC then further improves performance by adaptively coordinating global semantic information and local discriminative cues.

4.7 Efficiency-Performance Trade-off

To evaluate the practicality of QGF-Net, the efficiency-performance trade-off is reported in Table 10.

Table 10 Efficiency-performance trade-off.

Method	Params (M)	FLOPs (G)	Inference Time (ms/img)	FPS	Top-1 Accuracy (%)
Method A ¹	86.0	21.0	18.6	53.8	89.4
Method B ²	87.2	21.7	19.4	51.5	90.1
Method C ³	88.1	22.4	20.2	49.5	90.9
Method D ⁴	89.3	23.6	21.5	46.5	92.0

¹Method A = Self-Supervised ViT Baseline.

²Method B = + Classical Attention.

³Method C = + ICLCF + LDRP.

⁴Method D = QGF-Net.

QGF-Net achieves the highest accuracy while introducing only a moderate computational overhead compared with the self-supervised ViT baseline. The increase from 86.0M to 89.3M parameters and from 21.0G to 23.6G FLOPs is relatively modest considering the 2.6-point gain in Top-1 Accuracy. This indicates that QGF-Net maintains a reasonable balance between recognition performance and computational cost.

4.8 Training Stability under Different Quantum Settings

To further evaluate the practical behavior of the proposed quantum module, training stability under different quantum settings is summarized in Table 11.

The results show that the recommended configuration of 4 qubits, 2 layers, and reduced dimension 16 provides the best balance between mean performance, training stability, and convergence speed. Although the 6-qubit configuration reaches slightly higher mean accuracy, its larger variance and weaker stability make it less attractive as a default setting.

Table 11 Training stability under different quantum settings.

Qubits	Layers	Reduced Dim	Top-1 Mean (%)	Top-1 Std (%)	C.E. ⁵	Stability
2	1	8	90.8	0.18	52	1 ¹
4	1	16	91.4	0.14	47	2 ²
4	2	16	92.0	0.11	43	3 ³
6	2	32	92.2	0.31	49	4 ⁴

¹1 = Stable.²2 = Stable.³3 = Recommended.⁴4 = Slightly unstable.⁵C.E. = Convergence Epoch.

4.9 Visualization Analysis and Discussion

To further verify the rationality of QGF-Net at the mechanism level, visualization analysis is conducted from three perspectives, namely local region selection, attention response, and feature distribution.

First, in the visualization of local discriminative regions, the high-response regions selected by LDRP are mainly concentrated on semantically meaningful object parts such as the head, wings, tail, and local structural textures, while much less attention is paid to irrelevant background areas. Compared with the model without ICLCF, the full model exhibits more stable selected regions under different brightness and shadow conditions, indicating that ICLCF effectively alleviates the sensitivity of local features to illumination changes and improves the reliability of discriminative region selection.

Second, in the visualization of attention responses, QMA produces a more concentrated distribution over truly category-specific fine regions, while the classical attention model tends to assign more scattered responses to adjacent redundant areas. This observation is consistent with the quantitative gain achieved by QMA over classical attention in the experiments.

Finally, in t-SNE or UMAP feature embedding visualizations, QGF-Net learns a feature space with clearer inter-class separability and more compact intra-class clustering. Especially for highly similar subcategories, the incorporation of ICLCF, LDRP, and QMA leads to clearer decision boundaries between classes, indicating that QGF-Net not only improves classification accuracy, but also enhances the discriminative structure of the learned representation space.

4.10 Summary of This Section

This section has systematically evaluated QGF-Net on three representative benchmarks. The experimental results show that QGF-Net outperforms ResNet-50, standard ViT, the self-supervised ViT baseline, and the classical local attention variant in terms of Top-1 Accuracy, Macro-F1, and robustness. The ablation study further demonstrates that ICLCF, LDRP, QMA, and HQCC improve model performance from four

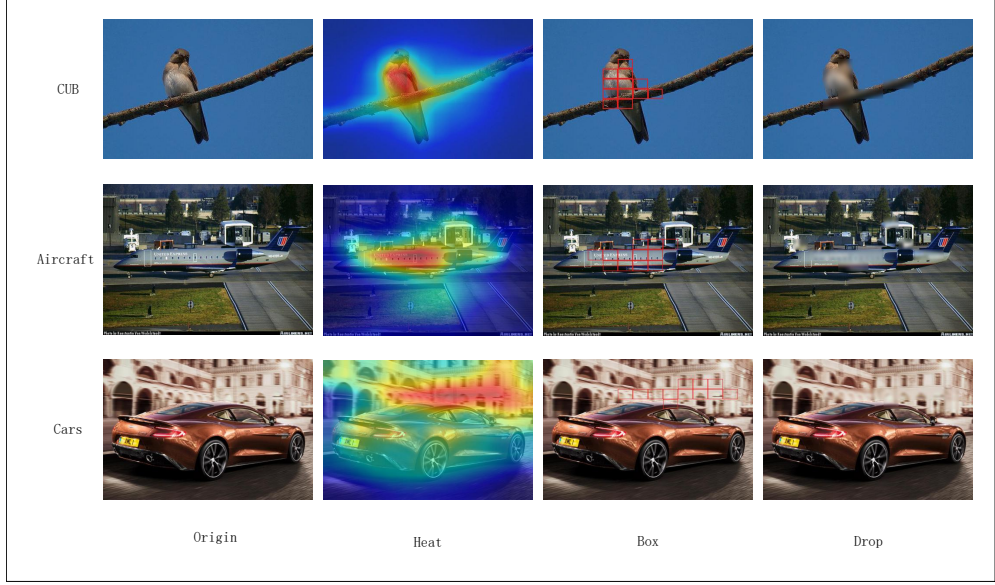


Fig. 5 Visualization of different branches in QGF-Net. The first column shows the original input image. The second column presents the heatmap of salient regions. The third and fourth columns show the annotated image and the attention-removed image, respectively. The results indicate that QGF-Net effectively captures the most discriminative regions for fine-grained classification, although some auxiliary informative regions may receive less attention

complementary perspectives, namely local stabilization, discriminative region compression, higher-order local relationship enhancement, and global-local collaborative classification.

5 Conclusion and Future Work

This paper focuses on three key challenges in fine-grained image classification: strong global semantic representation but insufficient local subtle-difference modeling, the sensitivity of local features to illumination changes and occlusion, and the limited capability of classical attention mechanisms in modeling higher-order local relationships [7, 10, 11]. To address these issues, this paper proposes QGF-Net, a hybrid quantum-classical fine-grained image classification network that integrates self-supervised visual representation learning, illumination-consistent local feature fusion, discriminative region selection, quantum measurement attention, and adaptive global-local classification into a unified framework.

QGF-Net adopts a DINOv2-pretrained Vision Transformer as the global semantic backbone, introduces ICLCF to perform illumination-consistent coarse fusion for local features, employs LDRP to explicitly select the most discriminative local regions, and further uses QMA to generate adaptive local region reweighting through quantum measurement, followed by HQCC for adaptive fusion and classification of global semantic and local discriminative information [11, 18–20].

Experimental results show that QGF-Net achieves 92.0%, 94.2%, and 89.5% Top-1 Accuracy on the CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets, respectively, outperforming ResNet-50, standard ViT, the self-supervised ViT baseline, and the classical local attention variant. Under perturbation settings including brightness variation, shadow simulation, and partial occlusion, QGF-Net also demonstrates stronger robustness, with an average performance drop of only 6.0%. The ablation studies further verify the effectiveness of ICLCF, LDRP, QMA, and HQCC, indicating that the performance gain of QGF-Net is not simply obtained by increasing model capacity, but rather by a modular design that progressively improves local feature stability, discriminative region selection, higher-order local relationship modeling, and global-local decision coordination.

From a methodological perspective, the core contribution of this work does not lie in merely introducing the concept of quantum computation into computer vision, but in constructing a complete problem-oriented pipeline for fine-grained recognition. Specifically, ICLCF first alleviates the instability of local features under illumination changes, LDRP then compresses a large number of patch tokens into a small set of highly informative discriminative regions, QMA subsequently enhances the complex relationships among these high-value local regions, and HQCC finally performs adaptive joint decision-making with both global semantic cues and local fine details. Therefore, the value of this work lies not only in its competitive classification performance, but also in providing an interpretable and extensible solution for further exploiting local discriminative information under a strong self-supervised backbone.

Nevertheless, this work still has several limitations. First, although the current experimental validation covers three representative fine-grained benchmarks, the cross-domain generalization of QGF-Net has not yet been systematically verified on larger-scale or long-tailed datasets such as NABirds. Second, although QMA outperforms classical attention in the experiments, it is still essentially a lightweight enhancement module based on small-scale parameterized quantum circuits in a compressed feature space, rather than a complete quantum vision model. Therefore, its advantage should be understood as a compact structural enhancement rather than a full replacement of classical visual learning paradigms [19, 20]. Third, because the official description of CUB-200-2011 explicitly notes a potential overlap with ImageNet/Flickr data, caution is still needed when interpreting the contribution of the proposed structure under large-scale pretrained backbones [14, 24].

Future work can be carried out in several directions. First, QGF-Net can be extended to more fine-grained datasets and cross-domain scenarios to verify its generalization ability under different category structures, imaging conditions, and sample scales. Second, the design of QMA can be further improved by introducing joint encoding of paired local regions, quantum correlation modeling across regions, or more efficient parameterized circuits in order to better characterize combinational relations among fine-grained local cues [18, 20]. Third, ICLCF can be integrated with stronger illumination-invariant learning strategies, self-distillation, or contrastive learning objectives to further improve the stability of local representations. Fourth, the method can be re-evaluated under stricter settings that better control pretraining data overlap, thereby more clearly separating the benefit of large-scale pretraining from

the gain brought by structural innovation. Finally, QGF-Net may also be explored in application scenarios that rely on subtle local differences, such as medical subtype recognition, industrial surface defect classification, and remote sensing fine-grained target recognition [2, 3].

In summary, this paper proposes a hybrid quantum-classical framework that integrates self-supervised visual representation, illumination-consistent local coarse fusion, discriminative region selection, quantum measurement attention, and adaptive global-local classification for fine-grained image classification. The results indicate that, even under a strong global visual backbone, further stabilizing, selecting, and enhancing high-value local regions remains a key path toward improving fine-grained recognition performance. At the same time, the quantum enhancement module shows promising potential as a compact supplementary component in such a framework [19, 20].

Acknowledgements. This work was supported by the Open Fund of the Advanced Cryptography and System Security Key Laboratory of Sichuan Province (Grant No. SKLACSS-202208).

Data availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Author contributions. YD conceived the experiment and provided the necessary financial support. XS conceived the experiment, conducted the experiments, and wrote the manuscript. SW and BW analysed the results. All authors reviewed the manuscript.

Declarations

Conflict of interest The author(s) declare no conflict of interest.

References

- [1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [2] Y. Xie, Q. Gong, X. Luan, J. Yan, and J. Zhang, “A survey of fine-grained visual categorization based on deep learning,” *Journal of Systems Engineering and Electronics*, no. 6, pp. 1337–1356, 2024.
- [3] J. M. Lim, K. M. Lim, C. P. Lee, and J. Y. Lim, “A review of few-shot fine-grained image classification,” *Expert Systems with Applications*, vol. 275, Art. no. 127054, 2025.
- [4] J. Fu, H. Zheng, and T. Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4438–4446.

- [5] H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning multi-attention convolutional neural network for fine-grained image recognition,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5209–5217.
- [6] X. Shi and B. Liu, “Boosting few-shot fine-grained image classification with global and local feature learning,” *The Journal of Supercomputing*, vol. 82, no. 2, p. 94, 2026.
- [7] H. Zhu, Y. Fu, J. Zhang, C. Wang, H. Lu, and S. Li, “Dual cross-attention learning for fine-grained visual categorization and object re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4691–4700.
- [8] X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, “Fine-grained object classification via self-supervised pose alignment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7398–7407.
- [9] D. Chang, Y. Tong, R. Du, Z. Ma, and T. Hospedales, “An erudite fine-grained visual classification model,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7268–7277.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
- [11] M. Oquab, T. Darcet, T. Moutakanni, et al., “DINOv2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
- [12] D. Kang, P. Koniusz, M. Cho, and N. Murray, “Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 19687–19696.
- [13] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200-2011 Dataset*. Pasadena, CA, USA: California Institute of Technology, 2011.
- [14] Perona Lab: Caltech-UCSD Birds-200-2011 (CUB-200-2011). https://www.vision.caltech.edu/datasets/cub_200_2011/. Accessed 23 Mar 2026
- [15] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2013, pp. 554–561.
- [16] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, “Fine-grained visual classification of aircraft,” *arXiv preprint arXiv:1306.5151*, 2013.

- [17] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.
- [18] D. Guala, S. Zhang, E. Cruz, C. A. Riofrío, J. Klepsch, and J. M. Arrazola, “Practical overview of image classification with tensor-network quantum circuits,” *Scientific Reports*, vol. 13, Art. no. 4427, 2023.
- [19] A. Senokosov, A. Sedykh, A. Saginalieva, B. Kyriacou, and A. Melnikov, “Quantum machine learning for image classification,” *Machine Learning: Science and Technology*, vol. 5, no. 1, Art. no. 015040, 2024.
- [20] E. A. Cherrat, I. Kerenidis, N. Mathur, J. Landman, M. Strahm, and Y. Y. Li, “Quantum vision transformers,” *Quantum*, vol. 8, p. 1265, 2024.
- [21] Y.-Q. Ling, J.-H. Zhang, L.-H. Zhang, Y.-R. Li, and H.-L. Huang, “Image classification using hybrid classical-quantum neural networks,” *International Journal of Theoretical Physics*, vol. 63, Art. no. 125, 2024.
- [22] M. A. Hafeez, A. Munir, and H. Ullah, “H-QNN: A hybrid quantum-classical neural network for improved binary image classification,” *AI*, vol. 5, pp. 1462–1481, 2024.
- [23] H. Asaoka and K. Kudo, “Quantum autoencoders for image classification,” *Quantum Machine Intelligence*, vol. 7, Art. no. 71, 2025.
- [24] X. Shi and B. Liu, “Boosting few-shot fine-grained image classification with global and local feature learning,” *The Journal of Supercomputing*, vol. 82, no. 2, p. 94, 2026.