Graph-Reasoning with Disentangled Attention: A Knowledge-Guided Approach for Fine-Grained Visual Classification

Yaosen Huang[1], Jiayao Xie[2], Juandong Huang[1], Tianrong Chen[1], Hua Tang[1,*]
[1]South China Normal University, Guangzhou, China
[2]Yunnan University, Kunming, China
{yaosenh, huangjuandong, 2023025160}@m.scnu.edu.cn,
464946673@qq.com, tanghua@m.scnu.edu.cn

***Abstract*—Fine-Grained Visual Classification confronts the dual challenge of distinguishing subtle inter-class differences and accommodating large intra-class variations. While existing methods address geometric and environmental variations effectively, they overlook Age-related Appearance Variation—drastic visual trait differences across life stages of the same species. To tackle this, we propose Graph-Reasoning with Disentangled Attention(GRDA), a unified framework integrating structural reasoning and explicit knowledge guidance. GRDA features a Disentangled Channel Attention module to decouple age-specific traits, and a Part-Aware Graph Attention Network that models invariant structural topology via semantic parts localized by Vision-Language Models. Additionally, we leverage Large Language Models to align visual features with stage-specific textual knowledge, bridging the semantic gap. We further curate the GuangDongBirds (GDB) dataset, which includes both adult and subadult instances. Extensive experiments show GRDA achieves competitive performance on GDB, CUB-200-2011, and NABirds, outperforming existing methods in handling complex biological variations.**

***Index Terms*—Fine-Grained Visual Classification, Graph Reasoning, Knowledge Guided Learning, Vision Language Model**

I. Introduction

Fine-Grained Visual Classification (FGVC) aims to recognize subordinate-level categories such as bird species. It is challenged by subtle inter-class differences and large intra-class variations, yet most existing methods only model geometric or environmental changes.

To advance research in this domain and address the lack of diverse, high-quality data for regional bird species, we introduce the GuangDongBirds (GDB) dataset. GDB is meticulously curated to cover a wide variety of bird species with high-resolution imagery. Uniquely, GDB captures birds across varying life stages, including a significant proportion of subadults individuals in addition to adults. This inclusion reveals a critical but often overlooked challenge in FGVC: Age-related Appearance Variation.

As illustrated in Figure 1, visual disparity between adults and subadults of the same species can exceed inter-species differences. For instance, the *Painted Bunting* in CUB-200-2011 [1] and *Nycticorax nycticorax* in GDB exhibit striking contrasts; adults and subadults often differ in coloration and pattern so significantly they resemble disjoint species. Standard FGVC methods often fail to reconcile these discrepancies, misclassifying subadults.

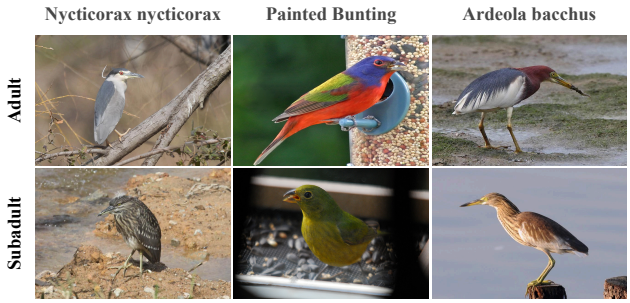

Fig. 1. Illustration of the Age-related Appearance Variation challenge. The visual disparity between adults and subadults is significant.

To address this, we integrate structural reasoning with knowledge guidance. We use Disentangled Channel Attention (DCA) to decouple age-specific traits, Part-Aware Graph Attention (PAGAT) to model part topology, and LLM-driven Knowledge Alignment to match visual features with age-specific textual descriptions.

Our contributions are summarized as follows:

- We identify and formalize the "Age-related Appearance Variation" problem in FGVC, supported by our newly curated GDB dataset which covers diverse life stages.
- We propose a unified framework combining DCA and PAGAT to robustly capture structural and part-level features invariant to age-related changes.
- We introduce an LLM-driven Knowledge Alignment mechanism that guides the visual model with explicit textual descriptions of adult and subadult characteristics, significantly improving robust recognition.

II. Related Works

To address the challenges of fine-grained classification, particularly the age-related appearance variations identified in

* Corresponding author.

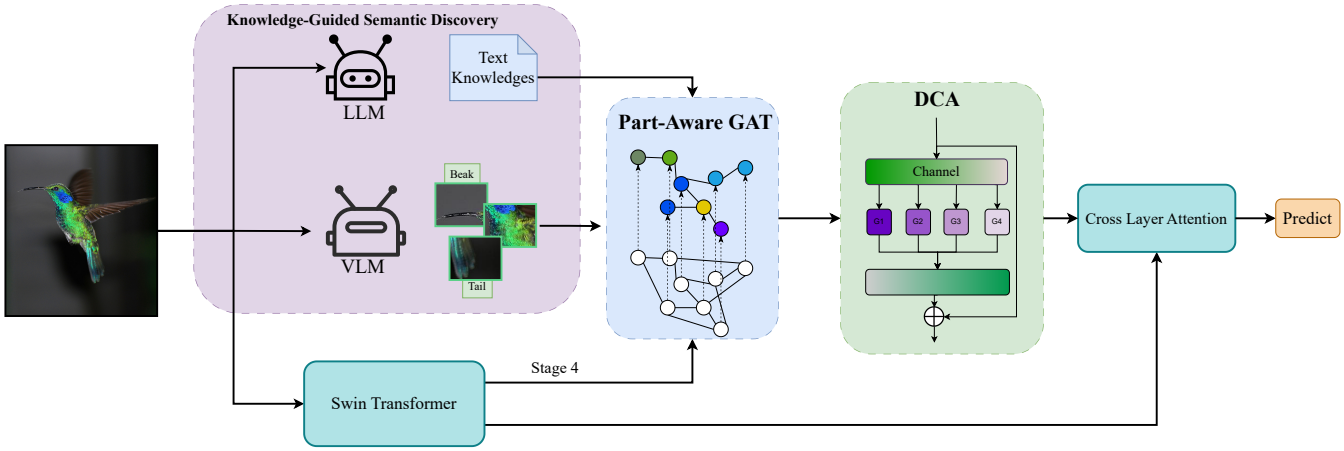

Fig. 2. Overview of the proposed framework. The Knowledge-Guided Semantic Discovery module utilizes LLM and VLM to provide text and part-level semantic guidance. These cues are integrated with Swin Transformer features via Part-Aware GAT (PAGAT), followed by refinement in the DCA module and aggregation through Cross Layer Attention.

this work, we review three lines of related research: modeling intra-class variance, graph reasoning for structural modeling, and knowledge-guided visual recognition.

A. Modeling Intra-Class Variation in FGVC

A central theme in FGVC is handling large intra-class variations while maintaining sensitivity to subtle inter-class differences. Existing methods largely focus on three specific types of variation:

a) Geometric Variations: Pose, viewpoint, and scale changes often drastically alter object appearance. Methods like P2P-Net [2] explicitly perform pose alignment to handle these variations, while RA-CNN [3] employs recurrent cropping to address scale differences.

b) Environmental Variations: Background clutter and illumination changes can distract models. Methods like CAP [4] use context-aware mechanisms to suppress background noise. Recent work GLEM [5] further integrates global and local features via attention erasure to reduce background interference and capture neglected details.

c) Local Deformations: Partial occlusions and missing parts are handled by approaches like TransFG [6] which aggregate cues from visible parts. GLSim [7] proposes a global-local similarity metric to improve robustness against such local variations by aligning feature patches.

However, these categories overlook a critical biological variation: life-stage progression. Unlike geometric or environmental changes, age-related variation changes the object's intrinsic texture and color patterns (e.g., juvenile plumage vs. adult plumage), requiring models that can look beyond surface appearance.

B. Graph Reasoning for Structural Modeling

Graph Convolutional Networks (GCNs) effectively model non-Euclidean relationships, capturing the invariant structural configuration of objects. SR-GNN [8] introduced a graph reasoning module to model inter-region dependencies based on feature similarity. Recently, methods like I2-HOFI [9] use GNNs to model high-order feature interactions through both inter- and intra-region graphs, seamlessly identifying global patterns and local details. Similarly, GKA [10] employs graph networks to mine adaptive connections between instances, enhancing feature discriminability.

In this work, we argue that while visual appearance varies across life stages, the structural topology of a bird remains constant. By employing a Graph Attention Network (GAT) within our PAGAT module, we explicitly model the relationships between semantic parts, providing a robust structural prior.

C. Knowledge-Guided Visual Recognition

Integrating external knowledge has become a key strategy in FGVC. While early methods relied on manual attributes, recent approaches leverage large models. Concept-Guided Learning (CGL) [11] and its extensions [12] utilize semantic concepts to guide representation learning. DOT-CBM [13] explores fine-grained visual-concept relations using optimal transport. With the rise of Large Vision-Language Models, methods like MCQA [14] transform zero-shot classification into a visual question-answering task, leveraging the comprehensive understanding of foundation models.

Our approach targets the specific "Age-related Appearance Variation" problem by utilizing LLMs to generate detailed textual contrasts between adult and subadult characteristics. By aligning our visual features with this explicit linguistic knowledge, our model learns to associate divergent visual appearances with the same underlying semantic category.

III. Method

The overall architecture, as illustrated in Figure 2, integrates a visual backbone with a knowledge-guided semantic discovery stream. The visual stream employs a Swin Transformer to extract hierarchical features. In parallel, the Knowledge-Guided Semantic Discovery module leverages Qwen3 [15]

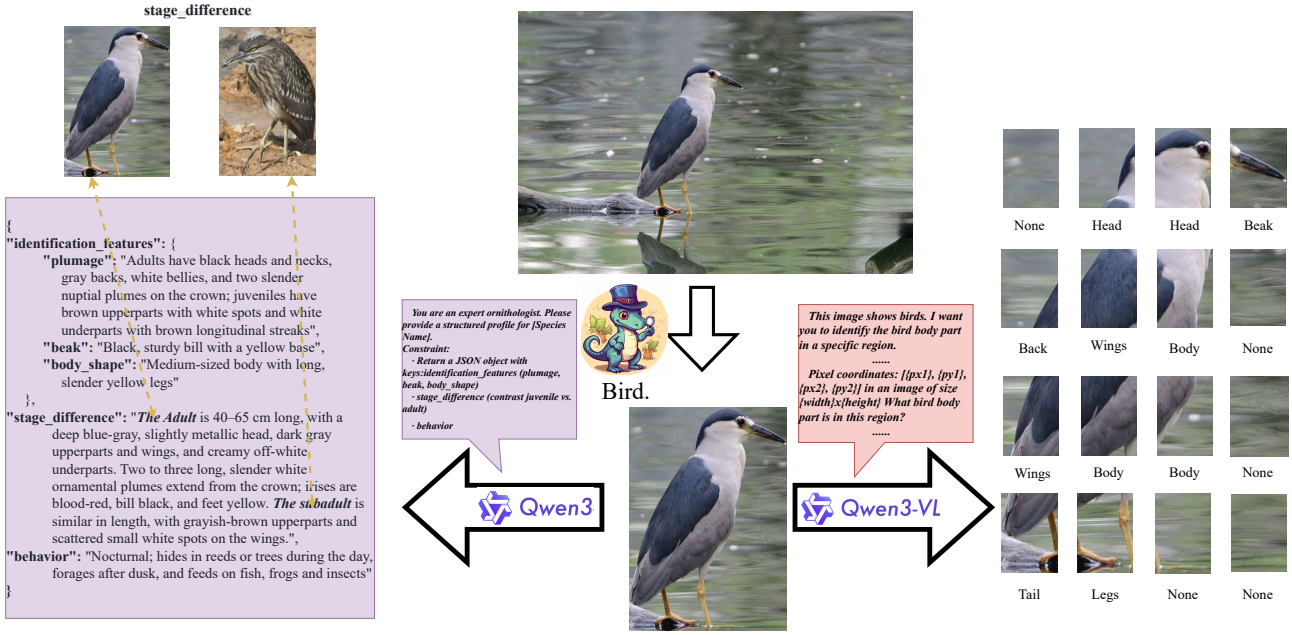

Fig. 3. **The Knowledge-Guided Semantic Discovery pipeline.** (Left) **LLM Stream:** Generates structured profiles with age-specific traits. (Right) **VLM Stream:** Identifies semantic parts via coordinate-based reasoning for graph construction.

for text knowledge extraction and Qwen3-VL [15] for visual part localization. These detailed semantic and part-level cues are fused with the visual features (Stage 4) via the PAGAT, which models the structural relationships between body parts. The output features are then refined by the DCA module to decouple diverse visual traits. Finally, the output of DCA and the multi-scale features from the Swin Transformer are aggregated through a Cross Layer Attention module [16] to produce the final classification prediction.

A. Knowledge-Guided Semantic Discovery

To bridge the semantic gap between visual appearance and biological identity, especially under the challenge of age-related variation, we construct a knowledge-guided pipeline (Figure 3) consisting of two streams: LLM-driven knowledge construction and VLM-assisted part localization, these components operate offline to generate external priors, ensuring computational efficiency during training.

a) LLM-Driven Stage-Aware Knowledge Construction.: We leverage LLMs to construct an offline, class-level knowledge base. To avoid data leakage, we generate a static profile for each category solely based on species names, serving as a fixed external dictionary. Specifically, we prompt Qwen3 to act as an ornithologist, creating structured profiles that explicitly decouple age-specific traits (e.g., distinguishing juvenile from adult characteristics) from general features. We then employ BERT [17] to encode these descriptions into semantic embeddings, which serve as invariant semantic prototypes during training. Crucially, during inference, the model relies solely on visual features aligned to this pre-computed semantic space without accessing ground-truth labels.

b) Detector-VLM Collaborative Part Localization.: To ground these semantics visually, we utilize a collaborative pipeline integrating open-set detection and VLMs for zero-shot part discovery. The process, illustrated in the right branch of Figure 3, involves two steps: 1) *Subject Grounding:* We first employ an open-set detector (Grounding DINO [18]) with the text prompt "Bird." to detect and crop the main subject, removing background noise such as water ripples or foliage. 2) *Semantic Part Reasoning:* The cropped image is uniformly partitioned into a 4×4 grid, yielding 16 distinct patches. Each patch is analyzed by Qwen3-VL to determine the enclosed body part. By aggregating these predictions, the VLM maps spatial grid regions to specific semantic parts (e.g., *Head, Beak, Wings, Body, Tail, Legs*) without manual supervision. These spatially-grounded semantic labels form the nodes $\mathcal{V}$ for our subsequent graph reasoning.

B. Part-Aware Graph Attention Network

a) Feature Sampling and Graph Construction: To model the invariant structural topology of birds despite age-related appearance variations, we propose a Part-Aware Graph Attention Network (PAGAT), as illustrated in Figure 4. We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each node $v_i \in \mathcal{V}$ $(i = 1 \dots N)$ represents a specific semantic body part (e.g., beak, wing). As shown in the "Sampling" process, we utilize the semantic grid assignments generated by the Detector-VLM pipeline to sample features directly from the Swin Transformer's output feature map. Specifically, for each region i, we extract a localized feature vector f_i^{vis} by aggregating features from the spatial grid regions associated with that part.

The connectivity $\mathcal{E}$ is defined by a dynamic adjacency matrix A, which integrates two robust priors: 1) Semantic

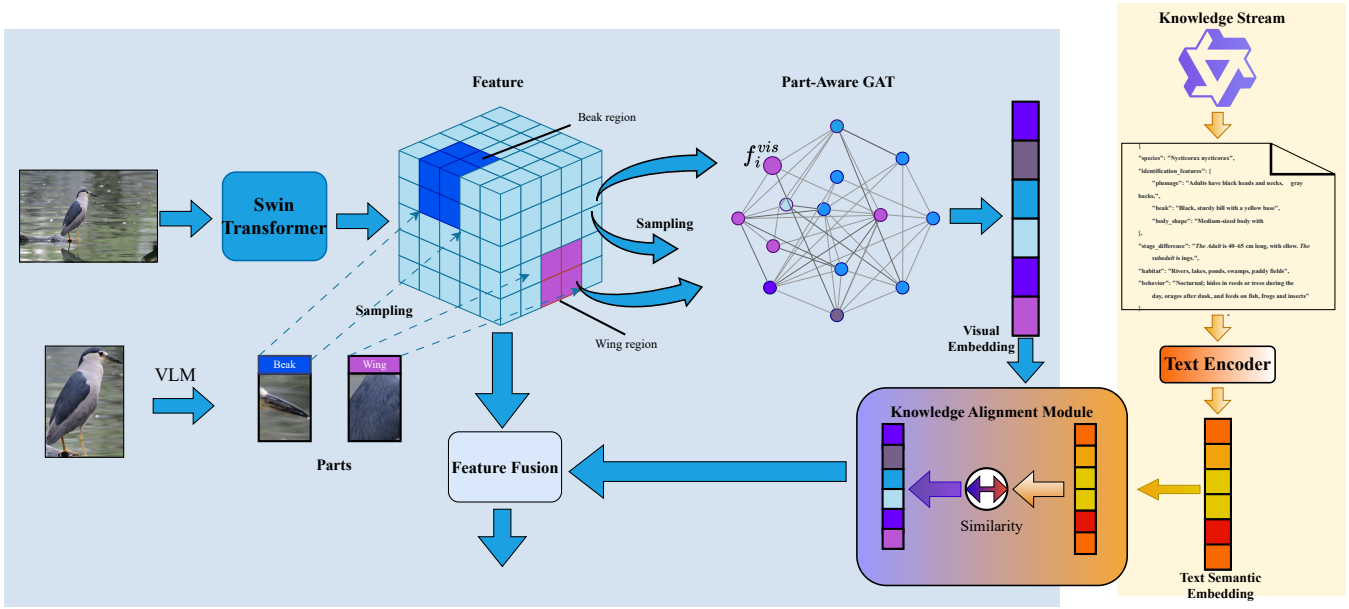

Fig. 4. **Part-Aware Graph Attention Networks.** It constructs a structural graph using semantic grid assignments and aligns visual features with textual knowledge embeddings encoded by a BERT-based text encoder.

Adjacency (A_{sem}): A learnable matrix initialized based on biological anatomical connections (e.g., head connects to neck), allowing the model to capture non-local dependencies. 2) Spatial Adjacency (A_{spa}): Derived from the normalized grid coordinates of parts, encoding the geometric configuration of the bird. The final adjacency is a weighted fusion: $A = \alpha A_{sem} + \beta A_{spa}$.

b) Part-Aware GAT construct: We employ the Part-Aware GAT to propagate information across the structural graph. For a node i, the updated feature h_i' is computed by aggregating messages from connected nodes weighted by the adjacency matrix A:

$$h_i' = \sigma\left(\sum_{j=1}^{N} A_{ij} W h_j\right), \tag{1}$$

where h_j denotes the feature of part j, W is a learnable weight matrix, and A_{ij} represents the connectivity strength between part i and part j (derived from the fused adjacency). This mechanism enables the model to enforce structural consistency. The resulting set of part features $F^{vis} = \{h_1', \ldots, h_N'\}$ constitutes the Visual Embedding.

c) Knowledge Alignment and Feature Fusion: As depicted in the Knowledge Stream, we employ a Text Encoder to transform the structured descriptions into a Text Semantic Embedding. The Knowledge Alignment Module then aligns the Visual Embedding with the Text Semantic Embedding to bridge the semantic gap. Furthermore, we introduce a Feature Fusion module to integrate the graph-reasoned structural features with the original spatial features. As shown in Figure 4, the Feature Fusion module takes the feature map from the Swin Transformer and the aligned Visual Embeddings as input. The fused features are then forwarded to the DCA Module for further refinement.

d) Optimization Objective: To ensure that the learned visual features align with biological knowledge, we introduce a Knowledge Alignment Loss ($\mathcal{L}_{align}$). Let $F^{text} = \{t_1, \ldots, t_N\}$ be the corresponding textual embeddings. We minimize the cosine distance between matched visual-textual pairs:

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{h_i' \cdot t_i}{\|h_i'\|\|t_i\|}\right). \tag{2}$$

Additionally, we impose a sparsity constraint $\mathcal{L}_{sparse}$ on the learned part importance weights to encourage the model to focus on the most discriminative regions.

C. Disentangled Channel Attention (DCA)

Standard channel attention mechanisms often treat all channels uniformly, potentially modifying global attributes while suppressing subtle fine-grained details. To address this, we introduce the Disentangled Channel Attention (DCA) module. As illustrated in Figure 5, DCA adopts a dual-branch architecture comprising a Local Context Branch and a Disentangled Global Branch.

The Local Context Branch captures local spatial patterns:

$$X_{local} = \text{PWConv}(\text{GELU}(\text{DWConv}(X))). \tag{3}$$

The Disentangled Global Branch partitions channels into G groups. For each group g, we compute attention to model distinct visual traits:

$$A_g = \text{Softmax}\left(\tau_g \cdot \frac{Q_g K_g^\top}{\sqrt{d_k}}\right) V_g, \tag{4}$$

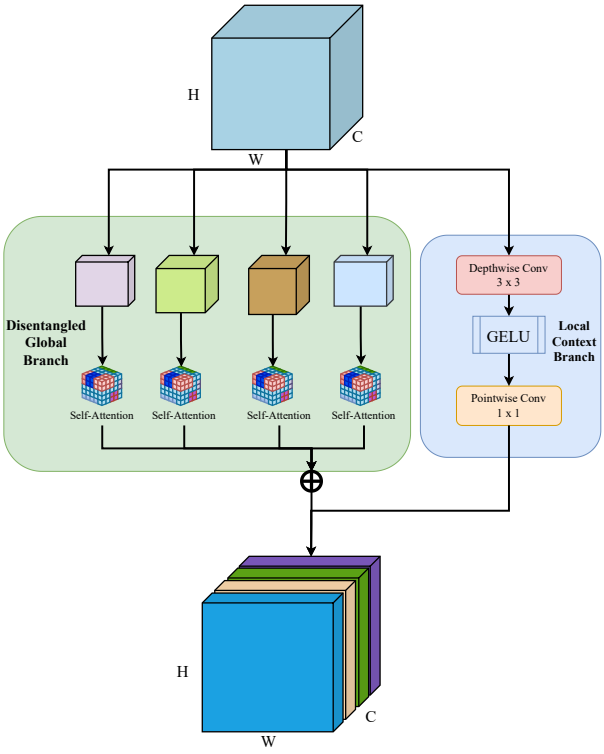

Fig. 5. **Disentangled Channel Attention (DCA).** It fuses a Local Context Branch (top) for spatial details and a Disentangled Global Branch (bottom) for modeling dependencies within distinct feature subspaces.

where Q_g, K_g, V_g are obtained via grouped linear projections, and τ_g is a learnable scale. The final output aggregates both branches:

$$X_{out} = \text{Proj}(\text{Concat}(\{A_g\}_{g=1}^{G}) + X_{local}). \quad (5)$$

This design allows our model to robustly handle age-related appearance variations by disentangling stage-specific features.

IV. Experiments

We conduct extensive experiments to validate the effectiveness of our proposed method across three fine-grained bird classification benchmarks. We compare our approach against recent state-of-the-art methods and perform comprehensive ablation studies to analyze the contribution of each component, demonstrating that our framework achieves competitive performance.

A. Datasets and Implementation Details

Datasets: We evaluate our method on three fine-grained bird classification benchmarks, whose statistics are summarized in Table I. GDB is our curated bird dataset comprising 20,539 images across 178 species. To address the long-tail distribution inherent in biological data, particularly for subadults, we adopt a stratified sampling strategy. The dataset contains 19,721 adult images and 818 subadult images, reflecting the natural scarcity of juvenile samples. We partition the subadult instances into 417 training and 401 test samples, creating a challenging few-shot scenario per category to rigorously evaluate the model's generalization across age-related variations. CUB-200-2011 [1] and NABirds [19] serve as supplementary benchmarks for additional validation.

TABLE I
STATISTICS OF THE FINE-GRAINED BIRD CLASSIFICATION BENCHMARKS.

Dataset	Species	Total	Training	Test
GDB	178	20,539	10,220	10,319
CUB-200-2011	200	11,788	5,994	5,794
NABirds	555	48,562	23,929	24,633

Implementation Details: We implement our method using PyTorch and conduct all experiments on a single NVIDIA RTX 3090 GPU. We use a pre-trained Swin Transformer Base (pre-trained on ImageNet-22K) as our backbone, with input images resized to 384×384 pixels.

B. Comparison with State-of-the-Art Methods

a) Classification Accuracy Evaluation: We compare our method against recent state-of-the-art approaches in fine-grained bird classification. Table II presents the comparative results on CUB-200-2011, NABirds, and our GDB dataset.

As shown in Table II, our method achieves 93.7% accuracy on GDB, outperforming the ViT-NeT by 0.9% and surpassing other methods. The improvement magnitude is larger on GDB compared to CUB-200-2011, as GDB contains a higher proportion of subadult instances while CUB has fewer. Despite this difference in data distribution, our method achieves consistent improvements on both benchmarks, demonstrating its effectiveness. GRDA also generalizes well, achieving 92.9% on CUB-200-2011 and 92.5% on NABirds.

b) Resource Consumption Evaluation: We evaluate the computational efficiency of our proposed method by comparing it with state-of-the-art Transformer-based approaches. Table III details the comparison in terms of parameters (Params), floating-point operations (FLOPs), and GPU memory consumption. Despite the introduction of the Part-Aware GAT and DCA modules, our method maintains a competitive efficiency profile. With an input resolution of 384×384, our model requires 48.8 GFLOPs, which is slightly lower than the baseline MPSA and significantly lower than methods like RAMS-Trans and ACC-ViT. Although the parameter count increases to 104.9M to accommodate the knowledge-guided components, the memory usage remains controlled at 1.36 GB, which is lower than ViT-Net. This indicates that our framework effectively balances high classification performance with reasonable resource consumption.

C. Performance on Age-Related Appearance Variations

To verify the effectiveness of our model in addressing Age-related Appearance Variations, we extracted the minority age group samples from the test set for species in the CUB and GDB datasets where distinct visual disparities exist between adults and subadults. Figure 6 presents a line chart that

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS. BEST RESULTS ARE **BOLD**, SECOND BEST ARE UNDERLINED.

Method	Publication	Backbone	CUB-200-2011	NABirds	GDB
Cross-X [20]	CVPR 2019	ResNet-50	87.7	86.2	91.3
GaRD [21]	CVPR 2021	ResNet-50	-	88.0	-
API-Net [22]	AAAI 2020	DenseNet	-	88.1	-
PMG-V2 [23]	ECCV 2020	ResNet-101	-	88.4	-
PART [24]	TIP 2021	ResNet-101	90.1	-	91.9
DATL [25]	ISVC 2020	Inception-v3	91.2	-	-
LGTF [26]	ICCV 2023	DenseNet	91.5	90.4	-
ABNT [27]	IJCNN 2024	ResNet-101	90.3	-	-
I2-HOFI [9]	IJCV 2025	Xception	91.6	**92.8**	-
ViT [28]	ICLR 2021	ViT-B/16	-	-	90.4
RAMS-Trans [29]	AAAI 2021	ViT-B/16	91.3	-	-
FFVT [30]	BMVC 2021	ViT-B/16	91.6	-	91.3
TransFG [6]	AAAI 2022	ViT-B/16	91.7	90.8	91.4
IELT [31]	TMM 2023	ViT-B/16	91.8	90.8	92.6
MP-FGVC [32]	AAAI 2024	ViT-B/16	91.8	91.0	-
ACC-ViT [33]	PR 2024	ViT-B/16	91.8	91.4	92.3
GLSim [7]	ISCAS 2025	ViT-B/16	91.1	92.0	91.6
GKA [10]	Neurocomputing 2025	ViT-B/16	91.8	-	-
CAMF [34]	SPL 2021	Swin-B	91.2	-	-
ViT-NeT [35]	ICML 2022	Swin-B	91.6	-	92.8
TransIFC [36]	TMM 2023	Swin-B	91.0	90.9	-
MPSA [16]	TIP 2024	Swin-B	92.8	92.4	92.8
CGL [12]	TIP 2025	Swin-T	92.6	91.7	-
CSQA-Net [37]	PR 2026	Swin-B	92.6	92.3	-
GRDA	-	**Swin-B**	**92.9**	92.5	**93.7**

TABLE III
COMPUTATIONAL COST ANALYSIS COMPARED WITH TRANSFORMER-BASED WORKS.

Backbone	Input Size	Params (M)	FLOPs (G)	Memory (GB)
ViT	448	86.4	78.5	1.54
IELT	448	93.5	73.2	1.18
ACC-ViT	448	87.0	162.9	2.01
Swin-Base	384	87.1	47.2	1.19
ViT-Net	448	92.2	65.6	1.41
TransIFC	384	87.4	52.4	1.43
MPSA	384	97.4	49.2	1.28
GRDA	384	104.9	48.8	1.36

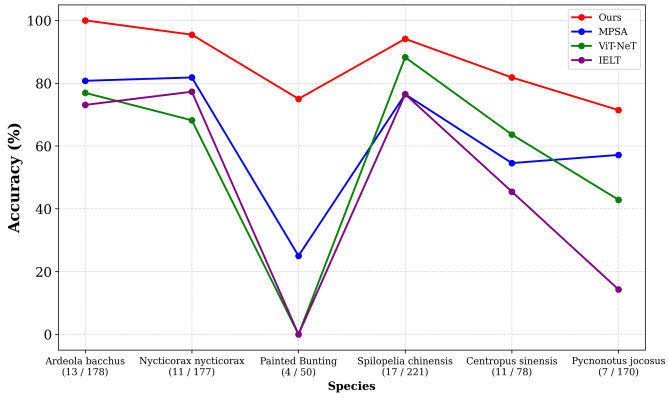

Fig. 6. Accuracy comparison on species with distinct age-related traits. The fraction below each species name indicates the sample count, where the numerator represents the minority age group (e.g., subadults or adults) in the test set and the denominator represents the total number of samples for that species.

compares the classification accuracy of our model with those of the baseline models MPSA and Swin-Transformer on these challenging species. The results show that our Knowledge-Guided approach achieves a significant performance improvement over the baselines in the accurate classification of subadult individuals, which demonstrates that our method is highly effective in handling the visual disparities between different life stages of the same species.

D. Visualization and Analysis

To intuitively demonstrate the effectiveness of our proposed method, we visualize the attention maps and part importance scores on representative images from the GDB dataset in Figure 7. As shown in the comparison between the second and third columns, our method produces heatmaps that are more accurately concentrated on the discriminative regions of the bird, such as the head and wings, whereas the baseline (MPSA) tends to have more scattered attention that may include background noise. This visual evidence suggests that the proposed Part-Aware GAT and DCA modules successfully guide the model to focus on the semantic regions crucial for robust fine-grained classification.

Additionally, the last column of Figure 7 displays the Part Importance scores corresponding to different semantic parts. To obtain these visualizations, we extract the learned importance weights from the Part-Aware GAT module and normalize them to the range [0, 1] for each sample via min-max normalization. These scores indicate the contribution of each part to the final decision. It can be observed that parts with high discriminative value, such as the head and

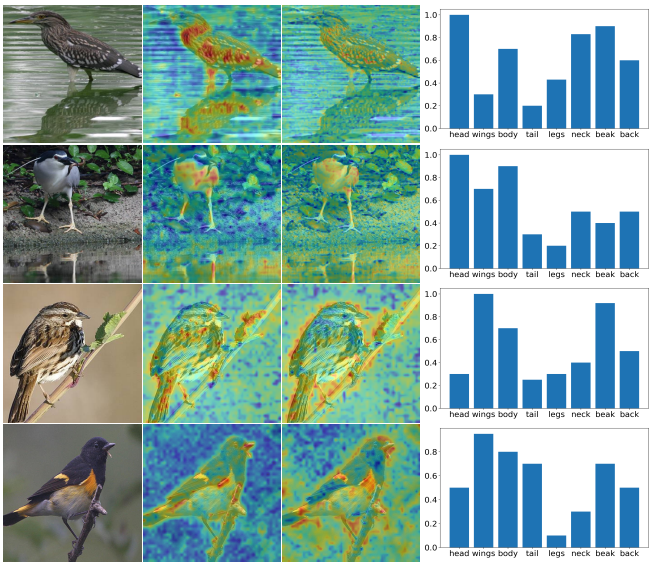

Fig. 7. Qualitative visualization on three bird datasets. Columns: Origin, Baseline, and GRDA(Ours).

wings, are consistently assigned higher importance scores. This distribution aligns with human expert knowledge in bird identification, validating that our knowledge-guided framework correctly leverages structural information to distinguish fine-grained categories.

E. Ablation Studies

To validate the contributions of the proposed components in GRDA—specifically DCA, PAGAT, and Knowledge Alignment—we conduct a series of ablation studies on the challenging GDB dataset. The results are summarized in Table IV.

TABLE IV
ABLATION STUDY OF KEY COMPONENTS IN GRDA.

Method	DCA	PAGAT	Knowledges	Acc.(%)
Baseline (MPSA [16])				92.80
w/ DCA	✓			93.12
w/ PAGAT		✓		93.30
w/ PAGAT + Knowledges		✓	✓	93.62
GRDA w/ Rand. Parts	✓	✓	✓	93.21
GRDA w/ Rand. Text	✓	✓	✓	93.33
GRDA	✓	✓	✓	**93.74**

The introduction of the DCA module (Row 2) improves the baseline performance from 92.80% to 93.12%, validating that disentangling channel features helps the model to better separate stage-specific visual traits from invariant features. Incorporating the Part-Aware GAT module (Row 3) further boosts accuracy to 93.30%, demonstrating the importance of explicit structural modeling to capture stable topological relationships. Further integrating LLM-driven knowledge into the Part-Aware GAT (Row 4) increases accuracy to 93.62%, highlighting the value of bridging the semantic gap with text embeddings. To analyze the sensitivity of our model to the accuracy of external knowledge and structural guidance, we conduct perturbation experiments with randomized inputs. Replacing precise part labels with random assignments (Row 5) drops performance to 93.21%, confirming that the topological graph relies on meaningful semantic regions. Crucially, substituting knowledge descriptions with random text (Row 6) simulates the scenario of maximum LLM hallucination or error. In this extreme case, accuracy reduces to 93.33%, which is lower than the full model (93.74%) but still significantly outperforms the baseline (92.80%). This indicates that while accurate biological information is essential for peak performance, our framework remains robust to potential knowledge extraction errors, relying on strong visual structural modeling as a fallback. Finally, combining all components in the full GRDA framework (Row 7) achieves the highest accuracy of 93.74%. This substantial overall improvement of 0.94% over the strong MPSA baseline confirms that the synergy between structural reasoning, knowledge guidance, and disentangled attention is critical for solving the Age-related Appearance Variation challenge.

F. Generalization Ability on CUB Dataset

To evaluate generalization, we integrate GRDA components into representative Swin and ViT-based methods on CUB-200-2011. For Swin Transformers, modules follow the final stage, while for ViT, patch tokens are reshaped to restore spatial structure for graph reasoning. As shown in Table V, our modules consistently boost performance.

TABLE V
IMPROVEMENTS BY PLUGGING GRDA COMPONENTS INTO OTHER METHODS ON CUB-200-2011.

Method	Backbone	Orig. (%)	+GRDA (%)	Impr. (%)
MPSA [16]	Swin-B	92.80	92.93	+0.13
ViT-NeT [35]	Swin-B	91.60	92.10	+0.50
TransFG [6]	ViT-B/16	91.70	91.81	+0.11
IELT [31]	ViT-B/16	91.80	91.90	+0.10
ACC-ViT [33]	ViT-B/16	91.80	91.87	+0.07

The modest gains on CUB are attributed to its limited subadult samples, yet the consistent improvements validate the backbone-agnostic effectiveness of our framework.

V. CONCLUSION

This paper addresses the under-explored Age-related Appearance Variation in Fine-Grained Visual Classification (FGVC). We propose GRDA, a unified framework integrating structural graph reasoning and knowledge-guided representation learning. Its Disentangled Channel Attention isolates stage-invariant features from age-induced variations. Incorporating LLM-driven knowledge and VLM-based semantic part discovery bridges the adult-subadult visual gap. Experiments on GDB, CUB-200-2011 and NABirds validate its superiority, highlighting the value of fusing structural priors with foundation model knowledge. Future work will focus on integrat-

ing habitat information via LLMs to enhance robustness for species with similar appearances but distinct niches.

ACKNOWLEDGMENT

This work was financially supported by the National Natural Science Foundation of China (82271267).

REFERENCES

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[2] X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, "Fine-grained object classification via self-supervised pose alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7399–7408.

[3] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.

[4] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Context-aware attentional pooling for fine-grained visual classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6356–6369, 2021.

[5] C. Zhou, D. Zhang, Q. He, M. Li, M. Li, and X. Zhou, "Glem: a global–local enhancement method for fine-grained image recognition with attention erasure and multi-view cropping," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 5, pp. 1–13, 2025.

[6] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "Transfg: A transformer architecture for fine-grained recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 852–860.

[7] E. A. Rios, M.-C. Hu, and B.-C. Lai, "Global-local similarity for efficient fine-grained image recognition with vision transformers," in *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2025, pp. 1–5.

[8] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Spatial relation graph neural network for fine-grained visual categorization," *IEEE Transactions on Image Processing*, vol. 30, pp. 4480–4493, 2021.

[9] A. Sikdar, Y. Liu, S. Kedarisetty, Y. Zhao, A. Ahmed, and A. Behera, "Interweaving insights: High-order feature interaction for fine-grained visual recognition," *International Journal of Computer Vision*, vol. 133, no. 4, pp. 1755–1779, 2025.

[10] Y. Wang, S. Ye, W. Hou, D. Xu, and X. You, "Gka: Graph-guided knowledge association for fine-grained visual categorization," *Neurocomputing*, vol. 634, p. 129819, 2025.

[11] J. Chen, J. He, S. Liu, J. Li, and Z. Luo, "Concept-guided learning for fine-grained visual categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3214–3228, 2023.

[12] Q. Bi, B. Zhou, W. Ji, and G.-S. Xia, "Universal fine-grained visual categorization by concept guided learning," *IEEE Transactions on Image Processing*, 2025.

[13] Y. Xie, Z. Zeng, H. Zhang, Y. Ding, Y. Wang, Z. Wang, B. Chen, and H. Liu, "Discovering fine-grained visual-concept relations by disentangled optimal transport concept bottleneck models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 199–30 209.

[14] M. Atabuzzaman, A. Zhang, and C. Thomas, "Zero-shot fine-grained image classification using large vision-language models," *arXiv preprint arXiv:2510.03903*, 2025.

[15] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[16] J. Wang, Q. Xu, B. Jiang, B. Luo, and J. Tang, "Multi-granularity part sampling attention for fine-grained visual classification," *IEEE Transactions on Image Processing*, 2024.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[18] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[19] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "The NaBirds dataset: Fine-grained bird recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. —.

[20] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, J. Li, J. Yang, and S.-N. Lim, "Cross-x learning for fine-grained visual categorization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8242–8251.

[21] Y. Zhao, K. Yan, F. Huang, and J. Li, "Graph-based high-order relation discovery for fine-grained recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 079–15 088.

[22] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 130–13 137.

[23] R. Du, J. Xie, Z. Ma, D. Chang, Y.-Z. Song, and J. Guo, "Progressive learning of category-consistent multi-granularity features for fine-grained visual classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9521–9535, 2021.

[24] Y. Zhao, J. Li, X. Chen, and Y. Tian, "Part-guided relational transformers for fine-grained visual recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 9470–9481, 2021.

[25] A. Imran and V. Athitsos, "Domain adaptive transfer learning on visual attention aware data augmentation for fine-grained visual categorization," in *International Symposium on Visual Computing*. Springer, 2020, pp. 53–65.

[26] L. Zhu, T. Chen, J. Yin, S. See, and J. Liu, "Learning gabor texture features for fine-grained recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1621–1631.

[27] B. Ding, X. Xu, X. Bao, N. Yan, and R. Zhang, "Abnt: Attention binary navigation tree for fine-grained visual classification," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] Y. Hu, X. Jin, Y. Zhang, H. Hong, J. Zhang, Y. He, and H. Xue, "Ramstrans: Recurrent attention multi-scale transformer for fine-grained image recognition," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4239–4248.

[30] Y. Li, Y. Li, H. Li, X. Zhang, W. Lin, and M. Zheng, "Ffvt: Feature fusion vision transformer for fine-grained visual categorization," *IEEE Transactions on Multimedia*, vol. 24, pp. 3825–3837, 2021.

[31] Q. Xu, J. Wang, B. Jiang, and B. Luo, "Fine-grained visual classification via internal ensemble learning transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 9015–9028, 2023.

[32] X. Jiang, H. Tang, J. Gao, X. Du, S. He, and Z. Li, "Delving into multimodal prompting for fine-grained visual classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 2570–2578.

[33] Z.-C. Zhang, Z.-D. Chen, Y. Wang, X. Luo, and X.-S. Xu, "A vision transformer for fine-grained classification by reducing noise and enhancing discriminative information," *Pattern Recognition*, vol. 145, p. 109979, 2024.

[34] Z. Miao, X. Zhao, J. Wang, Y. Li, and H. Li, "Complemental attention multi-feature fusion network for fine-grained classification," *IEEE Signal Processing Letters*, vol. 28, pp. 1983–1987, 2021.

[35] S. Kim, J. Nam, and B. C. Ko, "Vit-net: Interpretable vision transformers with neural tree decoder," in *International conference on machine learning*. PMLR, 2022, pp. 11 162–11 172.

[36] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, and Y.-F. Li, "Transifc: Invariant cues-aware feature concentration learning for efficient fine-grained bird image classification," *IEEE Transactions on Multimedia*, vol. 27, pp. 1677–1690, 2023.

[37] Q. Xu, S. Li, J. Wang, B. Jiang, B. Luo, and J. Tang, "Context-semantic quality awareness network for fine-grained visual categorization," *Pattern Recognition*, vol. 170, p. 112033, 2026.