

Dual-Interactive Coordinate Aggregation Network for Fine-Grained Visual Classification

Xuanyi Wu¹, Zixuan Yan², Jinling Wei^{1*}

¹School of Computer and Computing Science, Hangzhou City University, Hangzhou, China
2240101035@stu.hzcu.edu.cn, weijl@zucc.edu.cn

²College of Computer Science and Technology, Zhejiang University, Hangzhou, China
yanzixuan@zju.edu.cn

Abstract—Fine-grained visual categorization (FGVC) remains challenging due to subtle inter-class differences and large intra-class variations. Although Vision Transformer (ViT) has demonstrated strong global modeling capabilities, its patch-based representations often weaken local spatial awareness and restrict effective interaction across feature channels. To address these issues, we propose the Dual-Interactive Coordinate Aggregation Network to enhance feature representations. Specifically, we introduce a Coordinate-Aware Global Aggregation (CAGA) module that performs horizontal and vertical coordinate pooling to capture long-range spatial dependencies while preserving precise positional cues for accurate localization. Furthermore, to alleviate channel-wise information isolation, a Global Channel-Shuffle Attention (GCSA) module is designed by integrating grouped channel shuffling with depth-wise separable convolutions, facilitating dense global cross-channel interactions. Additionally, a multi-view feature fusion strategy is employed to jointly leverage enhanced local features and global semantic representations. Extensive experiments on CUB-200-2011, Stanford Cars, Stanford Dogs, and NABirds demonstrate that the proposed method consistently outperforms strong ViT baselines and achieves competitive performance against SOTA approaches.

Index Terms—Fine-Grained Visual Classification, Vision Transformer, Coordinate Attention, Channel Shuffling, Feature Interaction

I. INTRODUCTION

Fine-Grained Visual Categorization (FGVC) [1] has emerged as a fundamental task in computer vision, widely applied in scenarios ranging from ecological monitoring [2] and intelligent transportation [3] to smart retail [4] and smart agriculture [5], [6]. Unlike generic object recognition, FGVC aims to distinguish subordinate categories within a broader base class (e.g., bird species or vehicle models). This task is inherently challenging due to the “dual-variation” dilemma: subtle inter-class differences often reside in minute local details, while large intra-class variations caused by pose and background clutter obscure these discriminative features, as exemplified in Fig. 1.

Early approaches relied on hand-crafted features like SIFT or HOG [7], [8], which struggled with complex visual patterns. The introduction of Convolutional Neural Networks (CNNs), particularly Bilinear CNNs [9], significantly improved discriminability through enhanced feature interaction, though they remain limited by local receptive fields. Recently, Vision Transformers (ViTs) [10] have become the

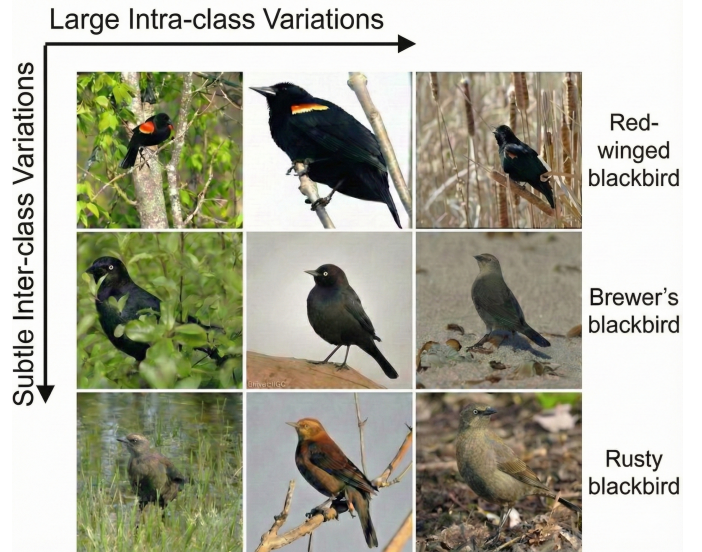


Fig. 1. Visualization of the “dual-variation” challenge on the CUB-200-2011 dataset. The vertical axis demonstrates subtle differences between similar species (Red-winged, Brewer’s, and Rusty Blackbirds), while the horizontal axis highlights dramatic appearance changes within the same class caused by environmental and pose factors.

dominant paradigm, utilizing self-attention for global modeling. Methods such as TransFG [11] and RAMS-Trans [12] have demonstrated ViT’s potential by selecting discriminative patches or employing multi-scale recurrent learning. However, the patch slicing process often leads to spatial information loss, and inconsistencies in multi-head attention can amplify background noise [13]. Existing methodologies generally fall into three categories: localization-based, feature fusion-based, and attention-enhancement methods. Localization approaches [14]–[16] attempt to identify salient regions but often lack dynamic global-local interaction or rely on inflexible bounding boxes. Feature fusion strategies [17], [18] integrate multi-view information but frequently incur high computational costs. Similarly, attention enhancement methods [19], [20] strengthen feature interaction but may not fully exploit the hierarchical nature of ViT features.

Despite these advancements, standard ViTs face three critical limitations in fine-grained tasks:

- global average pooling often discards spatial positional

information [21].

- insufficient interaction between feature channels allows background noise to interfere with semantic mining [22].
- over-reliance on the single [CLS] token neglects multi-view fusion [23].

These issues are particularly detrimental to fine-grained recognition, where discriminative cues are sparse, spatially localized, and often distributed across distant regions. To bridge these gaps, we propose the Dual-Interactive Coordinate Aggregation Network (DICA-Net), a novel architecture designed to synergize spatial coordinate encoding with dense channel interaction. Specifically, we introduce a Coordinate-Aware Global Aggregation (CAGA) module and a Global Channel-Shuffle Attention (GCSA) module. Unlike existing coordinate attention mechanisms designed for CNNs, CAGA is specifically tailored to ViT token representations and group-wise feature semantics. Complementing this, the GCSA module utilizes channel shuffling to break group barriers and facilitate deep feature interaction. Furthermore, a multi-view fusion strategy integrates the [CLS] token with globally pooled features for a comprehensive representation. Experiments demonstrate the effectiveness of our approach, relative to the ViT-B/16 baseline, our method yields improvements of 2.0%, 1.7%, 1.6%, and 1.2% on CUB-200, Stanford Cars, Stanford Dogs, and NABirds, respectively.

The main contributions of this paper are summarized as follows:

- We propose a Dual-Interactive Coordinate Aggregation Network (DICA-Net), which effectively overcomes the spatial limitations and insufficient channel communication of standard ViTs.
- We design the Coordinate-Aware Global Aggregation (CAGA) module to encode precise positional information and the Global Channel-Shuffle Attention (GCSA) module to foster deep cross-group feature interaction via channel shuffling, and construct a multi-view feature fusion strategy that integrates global semantics with background-suppressed and salient features.
- Extensive experiments demonstrate that our method achieves consistent improvements over strong ViT baselines across four benchmark datasets (CUB-200-2011, Stanford Cars, Stanford Dogs, and NABirds).

II. RELATED WORK

A. Attention Mechanism and Coordinate Encoding

Vision Transformer (ViT) [10] introduced the self-attention mechanism to computer vision, enabling the capture of long-range global dependencies that CNNs struggle to model with local receptive fields. However, the standard Multi-Head Self-Attention (MSA) in ViT suffers from quadratic computational complexity with respect to token length. This limitation is particularly challenging in Fine-Grained Visual Categorization (FGVC), where precise spatial localization is crucial as inter-subclass differences are often concentrated in minute local areas.

To address localization needs, attention mechanisms simulate the human visual system by dynamically weighting important regions. Early channel attention models like SE-Net [24] use global average pooling to compress spatial dimensions into 1×1 to model channel dependencies. Although effective, this operation inevitably discards spatial location information, making it difficult to pinpoint subtle key parts. To overcome these limitations, Coordinate Attention [25] embeds positional information into channel attention. Instead of 2D global pooling, it utilizes two 1D feature encoding operations to aggregate features along the horizontal (X) and vertical (Y) directions, respectively. This “factorization” strategy generates a pair of direction-aware feature maps, enabling the network to capture long-range dependencies—similar to ViT—but while preserving precise spatial coordinates with significantly lower computational cost. Compared to CBAM [26], this approach demonstrates superior localization capability.

B. Channel Shuffling and Feature Interaction

Beyond the spatial attention emphasized by ViT [10] and Coordinate Attention, effective information flow across feature channels is also critical for fine-grained recognition. Channel attention mechanisms, such as SE-Net [24] and ECA-Net [27], reweight channel responses using global pooling operations. While this improves discriminative representation, it often overlooks inter-channel interaction patterns induced by grouped convolutions.

To address channel-wise information isolation, ShuffleNet [28] introduced the Channel Shuffle operation, which enables cross-group information exchange with negligible computational overhead. This simple yet effective mechanism has been widely adopted in lightweight CNN architectures to enhance representation capability. However, most existing channel interaction and shuffling strategies are designed for convolutional feature maps and are rarely explored in Vision Transformer frameworks. This gap highlights the need for tailored mechanisms that can facilitate dense channel interaction while remaining compatible with Transformer-based representations.

III. METHOD

A. Overall Architecture

The overall architecture of DICA-Net is illustrated in Fig. 2. The network aims to enhance fine-grained feature representation by recovering spatial topology and fostering channel interaction, addressing the limitations of standard Vision Transformers. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we utilize a pre-trained ViT-Base [10] as the backbone. The image is split into N non-overlapping patches, flattened, and projected into a D -dimensional embedding space. A learnable class token ([CLS]) is prepended to the sequence. After passing through L Transformer Encoder layers, the output sequence is denoted as $Z_{seq} \in \mathbb{R}^{(N+1) \times D}$. To facilitate spatial-aware refinement, we extract the patch tokens $Z_{patch} \in \mathbb{R}^{N \times D}$ from the sequence and reshape them into a 2D feature map $X \in \mathbb{R}^{B \times D \times h \times w}$. This feature map is then sequentially

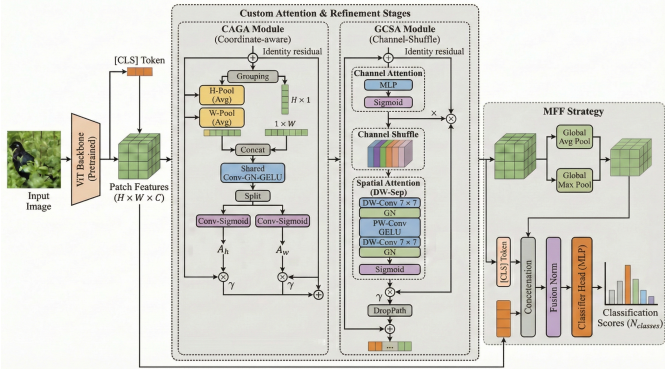


Fig. 2. DICA-Net Architecture Diagram

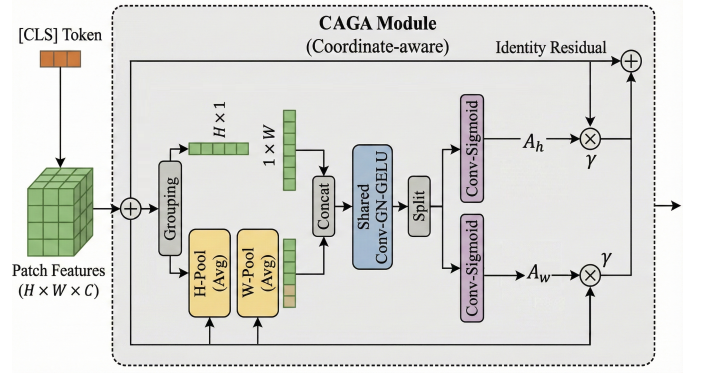


Fig. 3. CAGA Architecture Diagram

processed by the proposed CAGA and GCSA modules to reinforce spatial localization and channel semantic interaction. Finally, the enhanced features are aggregated via the MFF strategy for classification.

B. Coordinate-Aware Global Aggregation

In fine-grained classification, discriminative cues such as beak shapes or wing patterns are often spatially localized. However, standard Global Average Pooling compresses feature maps into a single vector, discarding spatial distribution information. To alleviate this issue, the CAGA module combines grouped convolutions with coordinate attention to explicitly encode positional cues while maintaining computational efficiency.

As shown in Fig. 3, the input feature map X is divided into G groups along the channel dimension, denoted as $[X_1, X_2, \dots, X_G]$, where each group $X_g \in \mathbb{R}^{C/G \times h \times w}$. To capture long-range dependencies with positional awareness, two 1D adaptive average pooling operations are applied along the height and width dimensions. For the c -th channel, the aggregated features are computed as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i), \quad z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (1)$$

The resulting direction-aware feature maps are concatenated and transformed by a shared 1×1 convolution F_1 , followed by Group Normalization (GN) and GELU activation δ , producing an intermediate feature f :

$$f = \delta(\text{GN}(F_1([z^h, z^w]))) \quad (2)$$

The feature f is then split into $f^h \in \mathbb{R}^{C/G \times h \times 1}$ and $f^w \in \mathbb{R}^{C/G \times 1 \times w}$. Two independent 1×1 convolutions F_h and F_w , followed by the Sigmoid function σ , generate the attention weights:

$$g^h = \sigma(F_h(f^h)), \quad g^w = \sigma(F_w(f^w)) \quad (3)$$

The final output of the CAGA module is obtained via a residual connection:

$$Y_{CAGA} = X + \gamma \cdot (X \odot g^h \odot g^w) \quad (4)$$

where γ is a learnable scaling parameter and $\odot$ denotes element-wise multiplication with broadcasting. By preserving grouped feature diversity and encoding coordinate information, CAGA enhances the localization of discriminative regions.

C. Global Channel-Shuffle Attention

Although the CAGA module enhances spatial localization, channel-wise semantic interaction remains insufficient due to grouped operations that isolate feature subspaces. To address this issue, we propose the Global Channel-Shuffle Attention (GCSA) module, which integrates channel attention, channel shuffling, and spatial attention, as shown in Fig. 4. Channel shuffling acts as a bridge between channel and spatial attention, enabling global feature interaction across channel groups. Given input features X , a channel attention branch based on a two-layer MLP with GELU activation is used to model channel dependencies:

$$M_c = \sigma(W_2(\delta(W_1(X)))) \quad (5)$$

The reweighted features are computed as $X' = X \odot M_c$. A Channel Shuffle operation is then applied to X' to promote information exchange across channel groups, which is the key component of the GCSA module. The shuffled features are further processed by a spatial attention branch with a hybrid convolutional bottleneck. A 7×7 depth-wise convolution captures long-range spatial dependencies, followed by a 1×1 point-wise convolution for channel fusion, and a standard 7×7 convolution for feature refinement. The spatial attention map is defined as:

$$M_s = \sigma(\text{GN}(\text{Conv}_{7 \times 7}(\delta(\text{PW}_{1 \times 1}(\text{GN}(\text{DW}_{7 \times 7}(X')))))) \quad (6)$$

The final output of the GCSA module is obtained using a residual connection with DropPath regularization:

$$Y_{GCSA} = X' + \text{DropPath}(\lambda \cdot X' \odot M_s) \quad (7)$$

where λ is a learnable scaling parameter initialized to zero for stable training.

D. Multi-View Feature Fusion (MFF)

To complement the global semantics of the [CLS] token Z_{cls} , MFF incorporates fine-grained local details from the refined features F . Specifically, we aggregate spatial information using Global Average Pooling (GAP) and Global Max Pooling (GMP) to capture background-suppressed and salient features. The final classification probability P is computed by fusing these views:

$$V_{fusion} = [Z_{cls}, \text{GAP}(F), \text{GMP}(F)] \quad (8)$$

$$P = \text{Softmax}(W_{head} \cdot \text{LN}(V_{fusion}) + b_{head}) \quad (9)$$

where $[\cdot]$ denotes concatenation. This fusion yields a robust representation $V_{fusion} \in \mathbb{R}^{3D}$ capable of handling large intra-class variations.

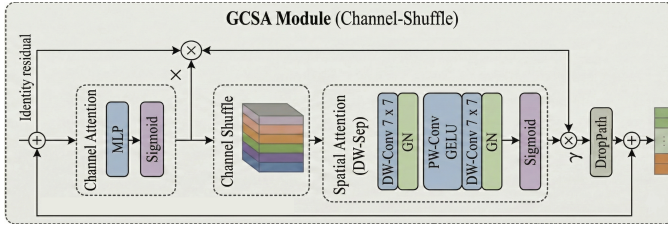


Fig. 4. GCSA Architecture Diagram

IV. EXPERIMENTS

A. Experimental Settings

1) *Datasets*: We conduct extensive experiments on four common fine-grained benchmarks, including CUB-200-2011 [29], Stanford Cars [30], Stanford Dogs [31] and NABirds [32]. In all experiments, we use the official train and validation splits for evaluation.

TABLE I
DATASET DISTRIBUTION

Dataset	Class	Train	Validation
CUB-200-2011	200	5994	5794
Stanford Cars	196	8144	8041
Stanford Dogs	120	12000	8580
NABirds	555	23929	24633

2) *Experimental Setup*: Experiments utilize a ViT-B/16 backbone pre-trained on ImageNet-21k, implemented on dual NVIDIA RTX 4090 GPUs. Input images are resized to 448×448 with augmentations including RandAugment and Random Erasing. The model is trained for 120 epochs using AdamW ($\text{lr}=3 \times 10^{-5}$, weight decay 0.05) with a cosine annealing scheduler and layer-wise learning rate decay (0.75). Model EMA and a fixed seed are used for stability.

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

TABLE II
ACCURACY(%) COMPARISON OF DIFFERENT METHODS ON FOUR DATASETS

Method	Backbone	CUB-200	S. Cars	S. Dogs	NABirds
ResNet [33]	ResNet-50	84.5	-	84.4	84.0
PMRC [34]	ResNet-50	91.8	-	93.1	84.0
API-Net [35]	Densenet-161	87.7	94.0	-	-
I2-HOFl [36]	ResNet-50	90.1	-	-	-
CAL [37]	ResNet-101	90.6	94.5	-	-
P2P-Net [38]	ResNet-50	90.2	94.4	-	-
ViT [10]	ViT-B/16	90.4	93.0	91.6	89.9
TransFG [11]	ViT-B/16	91.7	94.6	-	90.8
IELT [39]	ViT-B/16	91.8	94.2	91.8	90.8
R2-Trans [40]	ViT-B/16	91.5	93.8	92.8	90.2
Bi-ViT [41]	ViT-B/16	91.9	94.1	-	-
ACC-ViT [42]	ViT-B/16	91.8	-	92.9	91.0
DRMix [43]	ViT-B/16	91.8	-	92.9	91.0
FAL-ViT [44]	ViT-B/16	91.7	-	91.1	90.3
DICA-Net (Ours)	ViT-B/16	92.4	94.7	93.2	91.1

We compare the proposed DICA-Net with state-of-the-art methods, encompassing both CNN-based approaches (e.g.,

ResNet-50, API-Net) and Transformer-based approaches (e.g., TransFG, IELT). The quantitative results are summarized in Table II. DICA-Net demonstrates superior performance across all four benchmarks. On CUB-200-2011, DICA-Net achieves 92.4% Top-1 accuracy, outperforming the ResNet-50 baseline (84.5%) by 7.9% and consistently surpassing other ViT-based architectures. When compared to the ViT-B/16 baseline, our method delivers consistent absolute gains across all datasets: boosting accuracy by 2.0% on CUB-200, 1.7% on Stanford Cars, 1.6% on Stanford Dogs, and 1.2% on NABirds. Furthermore, DICA-Net outperforms recent competitive models such as TransFG and IELT. It is worth noting that on the challenging NABirds dataset, which contains 555 classes with fine-grained visual differences, our method reaches 91.1% accuracy. This result highlights the robustness of DICA-Net in handling large-scale categories with complex background interference.

C. Ablation Experiments

TABLE III
ACCURACY(%) COMPARISON WITH AND WITHOUT DIFFERENT MODULES ADDED TO THE BASELINE METHOD

CAGA	GCSA	MFF	CUB-200	S. Cars	S. Dogs	NABirds
-	-	-	90.45	93.00	91.52	89.90
-	-	✓	90.58	93.24	91.60	89.95
-	✓	-	91.45	93.75	92.05	90.25
✓	-	-	91.75	94.02	92.42	90.58
✓	✓	-	92.05	94.26	92.80	90.75
✓	✓	✓	92.42	94.65	93.18	91.12

1) *Effectiveness of Proposed Modules*: As shown in Table III, we analyze the impact of the Coordinate-Aware Global Aggregation (CAGA), Global Channel-Shuffle Attention (GCSA), and Multi-View Feature Fusion (MFF). On CUB-200-2011, integrating CAGA or GCSA independently improves accuracy to 91.75% and 91.45%, respectively, compared to the baseline (90.45%). This validates that CAGA effectively captures spatial position and GCSA enhances channel interaction. Combining CAGA and GCSA boosts accuracy to 92.05%, demonstrating that spatial anchoring and channel refining are complementary. Incorporating MFF further elevates performance to 92.42%. This 0.37% gain over the dual-module configuration confirms that fusing background-suppressed features (GAP) and salient features (GMP) with the [CLS] token retrieves rich semantic details ignored by the standard classification head.

2) *Hyperparameter Sensitivity*: We investigate two key hyperparameters: the group number G in CAGA and the kernel size k in GCSA.

a) *Grouping Number (G)*: Table IV shows that performance peaks at $G = 4$ (92.42%). Setting $G = 1$ (no grouping) degrades performance to 91.55%, as it fails to capture diverse feature patterns. Conversely, excessive grouping ($G = 16$) disrupts semantic continuity.

TABLE IV
EFFECT OF THE NUMBER OF GROUPS G IN THE CAGA MODULE ON $CUB - 200 - 2011$

Group Number (G)	Params (M)	FLOPs (G)	Top-1 Acc (%)
1	88.4	17.8	91.55
2	88.4	17.8	92.18
4 (Default)	88.4	17.8	92.42
8	88.4	17.8	92.26
16	88.4	17.8	91.93

b) Kernel Size (k): Table V indicates that increasing the convolution kernel size from 3×3 to 7×7 steadily improves accuracy (91.95% $\rightarrow$ 92.42%), confirming that larger receptive fields help associate discrete key parts. A further increase to 9×9 yields slight degradation while increasing computation, thus 7×7 is adopted as the default.

TABLE V
EFFECT OF THE KERNEL SIZE k IN THE GCSA MODULE ON $CUB - 200 - 2011$

Kernel Size ($k \times k$)	Top-1 Acc (%)	Inference Time (ms)
3×3	91.95	18.2
5×5	92.20	19.5
7×7 (Default)	92.42	21.1
9×9	92.38	24.6

3) Analysis of Fusion Strategies: We evaluate different pooling strategies for the MFF module in Table VI. Using only the [CLS] token yields 92.05%. Adding GAP improves this to 92.22% by suppressing noise. The combination of [CLS] + GAP + GMP achieves the best result (92.42%), proving that global max pooling provides complementary saliency information.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT TOKEN COMBINATIONS BASED ON THE DICA-NET ARCHITECTURE ON $CUB - 200 - 2011$

[CLS] Token	GAP (Avg)	GMP (Max)	Top-1 Acc (%)
Yes	No	No	92.05
Yes	Yes	No	92.22
Yes	No	Yes	92.15
Yes	Yes	Yes	92.42

4) Complexity Analysis: Table VII compares the model complexity. DICA-Net adds only 1.8M parameters (2%) and 0.2G FLOPs over the ViT-B/16 baseline. Compared to TransFG, our method achieves higher accuracy with an acceptable parameter increase, highlighting a favorable balance between performance and efficiency.

TABLE VII
COMPARISON OF MODEL PARAMETERS AND COMPUTATIONAL COMPLEXITY ON $CUB - 200 - 2011$

Method	Backbone	Params (M)	FLOPs (G)	Acc (%)
ResNet [33]	ResNet-50	25.6	4.1	84.5
ViT [10]	ViT-B/16	86.6	17.6	90.4
TransFG [11]	ViT-B/16	86.9	18.2	91.7
DICA-Net (Ours)	ViT-B/16	88.4	17.8	92.4

D. Attention Map Visualization

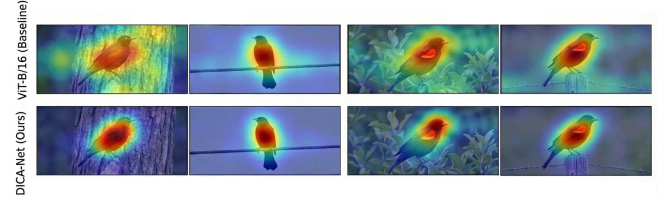


Fig. 5. Grad-CAM visualization results. Compared to the baseline (top), DICA-Net (bottom) demonstrates superior spatial focusing on key semantic parts and robustness against environmental clutter.

Fig. 5 visualizes the attention maps. The baseline ViT-B/16 (top row) tends to focus on the general object contour or is distracted by background noise (e.g., branches). In contrast, DICA-Net (bottom row) precisely localizes discriminative fine-grained regions, such as the specific shape of the beak or wing textures, verifying the effectiveness of the CAGA module in spatial anchoring.

V. CONCLUSION

This work systematically examines the limitations of Vision Transformers in fine-grained visual categorization, specifically addressing the degradation of spatial topology and restricted channel interaction. To overcome these challenges, we propose DICA-Net, a dual-interactive architecture that jointly enhances spatial awareness and channel-wise dependency. By integrating Coordinate-Aware Global Aggregation and Global Channel-Shuffle Attention, the proposed framework explicitly embeds positional cues and promotes dense cross-channel interaction. Comprehensive experiments on four FGVC benchmarks validate that DICA-Net achieves consistent improvements over state-of-the-art methods with minimal parameter overhead. These results demonstrate that restoring spatial-channel synergy is a principled and scalable direction for advancing high-granularity visual understanding.

ACKNOWLEDGMENTS

This work is partially supported by the Second Batch of Undergraduate Provincial Teaching Reform Project of the 14th Five-Year Plan of Zhejiang Province, China (JGBA2024676), and the Teaching Reform Project of Zhejiang Province, China (No. JGCG2025196).

REFERENCES

- [1] X.-S. Wei, Y.-Z. Song, O. Mac Aodha *et al.*, “Fine-grained image analysis with deep learning: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8927–8948, 2021.
- [2] Y. Liu, Y. Cai, Q. Jia *et al.*, “Novel class discovery for ultra-fine-grained visual categorization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 679–17 688.
- [3] S. D. Khan and H. Ullah, “A survey of advances in vision-based vehicle re-identification,” *Computer Vision and Image Understanding*, vol. 182, pp. 50–63, 2019.
- [4] X.-S. Wei, Q. Cui, L. Yang *et al.*, “Rpc: A large-scale and fine-grained retail product checkout dataset,” *Science China. Information Sciences*, vol. 65, no. 9, p. 197101, 2022.
- [5] L. Wei, K. Zhao, N. Wang *et al.*, “Fine-grained fish disease image recognition algorithm model,” *Laser & Optoelectronics Progress*, vol. 60, no. 16, p. 1610005, 2023.
- [6] G. Yu, G. Qingwen, W. Chuntao *et al.*, “Crop pest identification based on improved efficientnet model,” *Transactions of the Chinese Society of Agricultural Engineering*, vol. 38, no. 1, pp. 203–211, 2022.
- [7] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1. Ieee, 2005, pp. 886–893.
- [9] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear cnn models for fine-grained visual recognition,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [11] J. He, J.-N. Chen, S. Liu *et al.*, “Transfgr: A transformer architecture for fine-grained recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 852–860.
- [12] Y. Hu, X. Jin, Y. Zhang *et al.*, “Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition,” in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4239–4248.
- [13] A. Vaswani, N. Shazeer, N. Parmar *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [14] N. Zhang, J. Donahue, R. Girshick *et al.*, “Part-based r-cnns for fine-grained category detection,” in *European conference on computer vision*. Springer, 2014, pp. 834–849.
- [15] F. Zhang, M. Li, G. Zhai *et al.*, “Multi-branch and multi-scale attention learning for fine-grained visual categorization,” in *International Conference on Multimedia Modeling*. Springer, 2021, pp. 136–147.
- [16] L. Zhang, S. Huang, W. Liu *et al.*, “Learning a mixture of granularity-specific experts for fine-grained categorization,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8331–8340.
- [17] P. Zhao, S. Yang, W. Ding *et al.*, “Learning multi-scale attention network for fine-grained visual classification,” *Journal of Information and Intelligence*, 2025.
- [18] Y. Yao and E. Chen, “Care-net: A fine-grained image classification network via cross-layer semantic recalibration,” *International Journal for Housing Science & Its Applications*, vol. 47, no. 1, 2025.
- [19] Y. Wang, Y. Wang, Q. Ye *et al.*, “Strengthen contrastive semantic consistency for fine-grained image classification,” *Pattern Analysis and Applications*, vol. 28, no. 2, pp. 1–14, 2025.
- [20] Q. Xu, S. Li, J. Wang *et al.*, “Context-semantic quality awareness network for fine-grained visual categorization,” *Pattern Recognition*, vol. 170, p. 112033, 2026.
- [21] Z. Zhang, H. Tang, and J. Tang, “Multi-scale activation, selection, and aggregation: Exploring diverse cues for fine-grained bird recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 385–10 393.
- [22] Z.-G. Wang, S.-B. Chen, and B. Luo, “Background noise suppression for advanced fine-grained visual classification,” *Neurocomputing*, p. 132183, 2025.
- [23] D. Liu, L. Zhao, Y. Wang *et al.*, “Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification,” *Pattern Recognition*, vol. 140, p. 109550, 2023.
- [24] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [25] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.
- [26] S. Woo, J. Park, J.-Y. Lee *et al.*, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [27] Q. Wang, B. Wu, P. Zhu *et al.*, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.
- [28] X. Zhang, X. Zhou, M. Lin *et al.*, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.
- [29] C. Wah, S. Branson, P. Welinder *et al.*, “The caltech-ucsd birds-200-2011 dataset,” 2011.
- [30] J. Krause, M. Stark, J. Deng *et al.*, “3d object representations for fine-grained categorization,” in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.
- [31] A. Khosla, N. Jayadevaprakash, B. Yao *et al.*, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, no. 1, 2011.
- [32] G. Van Horn, S. Branson, R. Farrell *et al.*, “Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 595–604.
- [33] K. He, X. Zhang, S. Ren *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [34] K. Liu, K. Chen, and K. Jia, “Convolutional fine-grained classification with self-supervised target relation regularization,” *IEEE Transactions on Image Processing*, vol. 31, pp. 5570–5584, 2022.
- [35] 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.
- [36] A. Sikdar, Y. Liu, S. Kedarisetty *et al.*, “Interweaving insights: High-order feature interaction for fine-grained visual recognition,” *International Journal of Computer Vision*, vol. 133, no. 4, pp. 1755–1779, 2025.
- [37] Y. Rao, G. Chen, J. Lu *et al.*, “Counterfactual attention learning for fine-grained visual categorization and re-identification,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1025–1034.
- [38] X. Yang, Y. Wang, K. Chen *et al.*, “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.
- [39] Q. Xu, J. Wang, B. Jiang *et al.*, “Fine-grained visual classification via internal ensemble learning transformer,” *IEEE Transactions on Multimedia*, vol. 25, pp. 9015–9028, 2023.
- [40] S. Ye, S. Yu, Y. Wang *et al.*, “R2-trans: Fine-grained visual categorization with redundancy reduction,” *Image and Vision Computing*, vol. 143, p. 104923, 2024.
- [41] Y. Li, S. Xu, M. Lin *et al.*, “Bi-vit: Pushing the limit of vision transformer quantization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3243–3251.
- [42] N. Ibtchaz, N. Yan, M. Mortazavi *et al.*, “Acc-vit: Atrous convolution’s comeback in vision transformers,” *arXiv preprint arXiv:2403.04200*, 2024.
- [43] S. Wang, Z. Wang, Y. Chen *et al.*, “Drmix: Decomposition-recomposition data augmentation with diffusion model,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 7481–7489.
- [44] Y. Huang, Z. Hechen, M. Zhou *et al.*, “An attention-locating algorithm for eliminating background effects in fine-grained visual classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.