Multi-scale Activation, Refinement, and Aggregation: Exploring Diverse Cues for Fine-Grained Bird Recognition

Zhicheng Zhang[1], **Hao Tang**[1,2], **Jinhui Tang**[1]*

[1] Nanjing University of Science and Technology, China
[2] Centre for Smart Health, The Hong Kong Polytechnic University
{zzc_666, jinhuitang}@njust.edu.cn, howard.haotang@gmail.com

Abstract

Given the critical role of birds in ecosystems, Fine-Grained Bird Recognition (FGBR) has gained increasing attention, particularly in distinguishing birds within similar subcategories. Although Vision Transformer (ViT)-based methods often outperform Convolutional Neural Network (CNN)-based methods in FGBR, recent studies reveal that the limited receptive field of plain ViT model hinders representational richness and makes them vulnerable to scale variance. Thus, enhancing the multi-scale capabilities of existing ViT-based models to overcome this bottleneck in FGBR is a worthwhile pursuit. In this paper, we propose a novel framework for FGBR, namely Multi-scale Diverse Cues Modeling (MDCM), which explores diverse cues at different scales across various stages of a multi-scale Vision Transformer (MS-ViT) in an "Activation-Selection-Aggregation" paradigm. Specifically, we first propose a multi-scale cue activation module to ensure the discriminative cues learned at different stage are mutually different. Subsequently, a multi-scale token selection mechanism is proposed to remove redundant noise and highlight discriminative, scale-specific cues at each stage. Finally, the selected tokens from each stage are independently utilized for bird recognition, and the recognition results from multiple stages are adaptively fused through a multi-scale dynamic aggregation mechanism for final model decisions. Both qualitative and quantitative results demonstrate the effectiveness of our proposed MDCM, which outperforms CNN- and ViT-based models on several widely-used FGBR benchmarks.

Introduction

Birds play a crucial role in the biological chain. However, the advancement of industrialization and consequent environmental degradation have led to a sharp decline in numerous bird species. Statistics (Malhotra 2022) indicate that nearly half of all bird species are in decline. Due to their sensitivity to environmental changes, birds' activities can serve as indicators of environmental changes. Fine-Grained Bird Recognition (FGBR) has become a research hotspot, crucial for the conservation of bird species and natural habitats.

Fine-grained bird recognition is widely recognized as a challenging task due to the intrinsic characteristics of bird images (Tang et al. 2022; Jiang, Tang, and Li 2024). These challenges can be summarized as follows:

- **Complex Backgrounds:** Bird habitats, whether in the wild or urban environments, often characterized by complex and variable backgrounds that can obscure the subtle features distinguishing one species from another. For example, as shown in Figure 1, the key difference between the Yellow-Throated Vireo and the White-Eyed Vireo lies in the eye color. Complex backgrounds can misdirect the model's attention, causing it to focus on irrelevant background elements instead of critical features, leading to potential misclassification.
- **Scale Variations:** Bird images often exhibit significant scale variations due to differences in shooting dis-

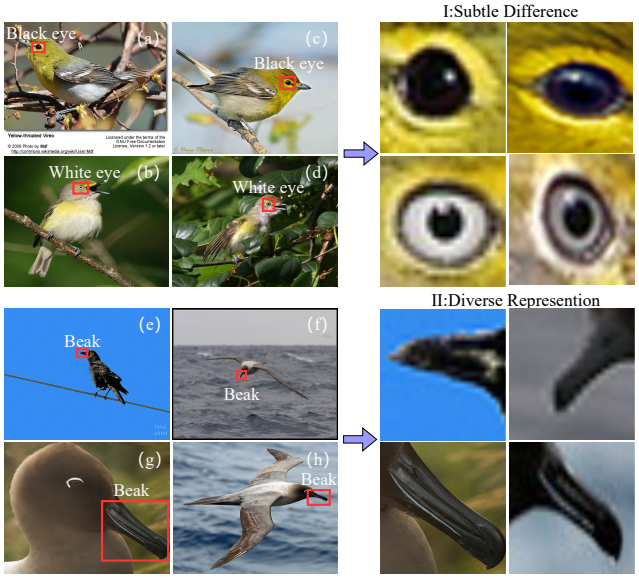

Figure 1: The primary challenges in FGBR are evident from bird images. Figures (a) to (d) show subtle species differences, often obscured by complex backgrounds. Figures (e) to (h) highlight significant scale variations between distant and close-up shots, making the same body parts appear different and complicating the recognition task.

*Corresponding author.

tances, resulting in varying sizes of the same body part (*e.g.,* beak and eyes) across images (Figure 1). These multi-scale differences complicate accurate recognition, requiring the model to consistently identify features such as the shape and texture of the beak or the color of the eyes, regardless of the scale at which they appear.

Recently, Vision Transformers (ViTs) (Dosovitskiy et al. 2021) have demonstrated impressive performance across various visual tasks (Jiang et al. 2024b; Shen et al. 2023). ViT divides images into patches, projects these patches into tokens, and then feeds them into transformer blocks for image modeling. However, pure ViT architectures face challenges when applied to fine-grained recognition tasks (Fang et al. 2024; Zha et al. 2023). Firstly, ViT models often struggle to capture the discriminative body parts crucial for birds. To address this issue, researchers have developed methods that encourage the model to focus on object regions rather than background noise. These approaches include selecting specific image patches that contain relevant features (He et al. 2022a; Wang, Yu, and Gao 2021) or designing mechanisms that emphasize discriminative cues (Liu et al. 2023; Zhu et al. 2022). Secondly, ViT's reliance on fixed-size image patches makes it less effective at capturing multi-scale features. To overcome this limitation, previous works have introduced multi-scale modules, which allow the model to observe images at different scales across multiple stages and use the features extracted from the final stage for downstream tasks (Zhang et al. 2021; Liu et al. 2021; Li et al. 2022; Tang et al. 2023; Dong et al. 2024b). However, significant scale variations in FGBR remain difficult to manage, as certain features learned at appropriate scales in earlier stages may be discarded during later computations, leading to the loss of valuable information.

Unlike previous works on FGBR that primarily exploit general bird features at a single scale, this work proposes a novel "Activation-Selection-Aggregation" paradigm. This approach captures diverse cues by extracting multi-scale features across various stages of a multi-scale Vision Transformer (MS-ViT). To obtain these diverse cues, we first extract features from multiple stages and introduce a multi-scale cue activation module during the forward process to adjust and enhance cue distinctiveness. Consistent with this architecture, we then design the multi-scale token selection mechanism, which leverages deep semantic information to guide token selection in shallow layers, thereby enhancing robustness. With cues learned across multiple stages, we could achieve recognition results at various scales. However, due to significant scale variations in FGBR, recognition results at inappropriate scales can negatively impact overall accuracy. To address this, we implement a gating mechanism that dynamically aggregates the recognition results. We validate the effectiveness of our method on two popular bird datasets and a large-scale species recognition dataset. Additionally, we also provide intuitive visualizations demonstrating how our approach mitigates background noise and offer quantitative analyses of the multi-scale cues aggregation's performance.

Related Work

The primary challenges of FGBR include high visual similarity among subclasses, complex and variable background environments, and multi-scale representations. Naturally, researchers have explored part localization to extract part-level features, performing feature interaction and alignment for classification. Early approaches typically designed localization sub-networks for FGBR (Zhang et al. 2014; Lin et al. 2015; Zhang et al. 2016), but these methods often required highly accurate part-level annotations for training, which limits its development. In response, later researchers shifted towards weakly supervised methods, such as Region Proposal Networks (RPN) and attention mechanisms (Ge, Lin, and Yu 2019; Tang et al. 2020; He et al. 2022a; Ke et al. 2023; Yan et al. 2023; Dong et al. 2024a).

However, the localization process significantly increases computational cost. Given the crucial role of feature learning in visual tasks, researchers have sought to enhance feature learning to capture subtle differences. (Shen et al. 2022) introduces additional convolutional layers to assess the importance of different regions within the feature map, adjusting activation relationships to uncover rich cues. Similarly, (Song and Yang 2021) employs extra convolutional layers to determine the importance of the feature map, adjusts its activations to encourage the model to focus on more potential regions. Some other researchers focus on identifying the most salient features and leveraging them for interaction and recognition. (He et al. 2022a) selects the most important tokens before the last layer to encourage the model to focus on discriminative parts. (Wang, Yu, and Gao 2021) aggregates the important tokens from each transformer layer to capture discriminative features. Similar approaches have been explored by (Hu et al. 2021; Chou, Kao, and Lin 2023). Additionally, some methods have investigated pairwise feature interactions to derive a unified yet discriminative image representation from paired images of different species (Zhuang, Wang, and Qiao 2020; Zhu et al. 2022).

Multi-scale Diverse Cues Modeling

Figure 2 illustrates the overall framework of our MDCM, which follows an "Activation-Selection-Aggregation" paradigm. Firstly, the feature extractor is employed to capture cues at different scales across various stages. During the forward pass, we adjust the activation strength of these features to ensure that each stage learns distinct features. Secondly, to better focus on the salient parts of birds, we remove irrelevant background patches, highlighting regions crucial for classification. Finally, the multi-scale feature representations are fed into multiple classifiers to generate classification results, which are then aggregated through a gating mechanism for precise target classification.

Multi-Scale Vision Transformer

We utilize Multi-scale Vision Transformer (MS-ViT) (Li et al. 2022) as our primary backbone, due to its ability to capture multi-scale features from images effectively. Unlike traditional Vision Transformer (Dosovitskiy et al. 2021), which divides image into patches and project these

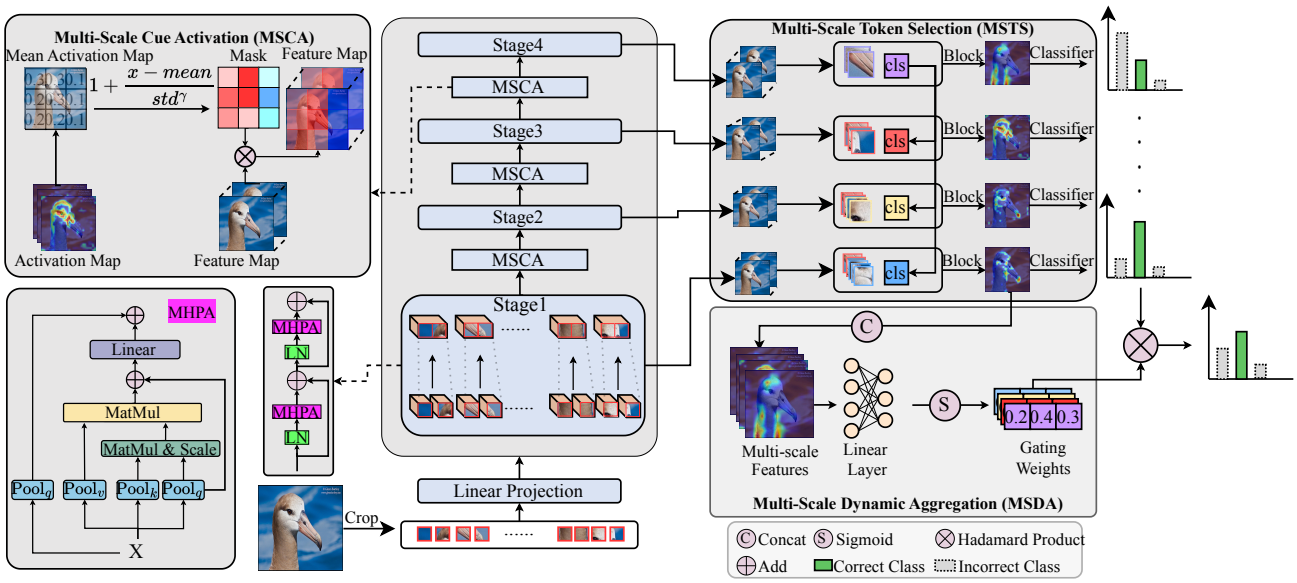

Figure 2: The framework of our MDCM. During the forward pass, MSCA adjusts the activation of feature map to ensure the cues learned at different stage are mutually different. Subsequently, MSTS extracts diverse cues from multiple stages and filters out noisy regions. Finally, MSDA dynamic aggregates the classification results for the final model decisions.

patches into tokens, MS-ViT introduces a pooling operation within the multi-head self-attention mechanism, termed Multi-Head Pooling Attention (MHPA). This approach reduces the number of tokens, increases the number of channels, lowers the image resolution, and adjusts the scale.

Let $X_0 \in \mathbb{R}^{h_0 \times w_0 \times c_0}$ represents the bird image, where h_0, w_0, and c_0 denote the height, width, and channels of the input image, respectively. Initially, the input image is divided into patches of size $o \times o$, which are then reshaped into one-dimensional patches sequence and passed through the patch embedding layer. A trainable class (`cls`) token is appended to the patch sequence, and a learnable position embedding is added to retain positional information. The embedding process is expressed as follows:

$$\mathbf{Z} = [z^0, z^1\mathbf{E}, z^2\mathbf{E}, \cdots, z^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{(\frac{h_0 \cdot w_0}{o^2} \cdot c_0) \times D}$ is the patch embedding projection, $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \cdot D}$ represents the position embedding, N denotes the number of image patches, z^0 is the learnable `cls` token, and z^n for $n \in \{1, 2, \ldots, N\}$ are the image patches. The MS-ViT consists of L layers of Multi Head Pooling Attention (MHPA) and Multi-Layer Perceptron (MLP) blocks. The forward pass for the l-th layer is as follows:

$$\mathbf{Z}'_l = \text{MHPA}(\text{LN}(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1}, \quad l \in 1, 2, \ldots, L, \tag{2}$$

$$\mathbf{Z}_l = \text{MLP}(\text{LN}(\mathbf{Z}'_l)) + \mathbf{Z}'_l, \quad l \in 1, 2, \ldots, L, \tag{3}$$

where $\text{LN}(\cdot)$ denotes the layer normalization operation. MS-ViT utilizes the `cls` token of the last encoder layer, z^0_L, as the representation of the global feature, which is then forwarded to a classifier head for the final recognition.

Multi-Scale Cue Activation

If cues learned at different stages of MS-ViT are highly similar, it would impede the effective utilization of multiscale information in bird images. Inspired by the success of the attention-based erasing mechanism in CNN-based architectures (Song and Yang 2021; Shen et al. 2022), we propose the parameter-free Multi-Scale Cue Activation (MSCA) module for MS-ViT. This module adjusts the activation of discriminative cues at each stage to explicitly learn multiple scale-specific representations, distinguishing it significantly from previous ViT-based methods.

During the forward propagation of the MS-ViT, the image is encoded into tokens for computation. Since the activation strength of cues should not be adjusted arbitrarily, it is crucial to first assess the importance of each token. Specifically, in MHPA, the tokens $z \in \mathbb{R}^{(N+1) \times D}$ are mapped and pooled into three matrices: $\mathbf{Q}$ (query), $\mathbf{K}$ (key), and $\mathbf{V}$ (value). The attention operation is performed as $\text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}$, where d is the dimension of the query vector. It is widely known that the `cls` token in ViTs tends to focus more on class-specific tokens rather than on tokens associated with non-object regions. Therefore, we propose utilizing the attentiveness of the `cls` token to other tokens as an activation map to identify the most important cues. Specifically, the activation map $\mathbf{A}$ is defined as $\mathbf{A} = \text{Softmax}(\frac{\mathbf{q}_{class} \cdot \mathbf{K}^\top}{\sqrt{d}})$, where $\mathbf{q}_{class}$ represents the query vector of the `cls` token, and $\mathbf{K}$ and $\mathbf{V}$ correspond to the key and value matrices, respectively. Consequently, the activation value a_i (*i.e.*, the i-th entry in $\mathbf{A}$) reflects the importance of the i-th token.

Some studies (Serrano and Smith 2019; Abnar and Zuidema 2020) indicate that high-level attention informa-

tion may not always align with the true importance of the input tokens. To enhance robustness, we average the activation maps from previous stages to form the attention map for the current stage. To address the shape mismatch caused by pooling operations, we employ interpolation to adjust the shapes accordingly. Subsequently, we map the importance scores to an enhancement and suppression mask to adjust the activation of cues without introducing additional parameters. Specifically, for each element $a^k \in \{a^1, a^2, \cdots, a^N\}$ in $\mathbf{A}$, we apply the following operation to derive the scaling mask:

$$m^k = \frac{a^k - \text{mean}(\mathbf{A}')}{\text{std}(\mathbf{A}')^\gamma}, \tag{4}$$

where $\mathbf{A}' \in \mathbb{R}^{N_k}$ represents all elements in $\mathbf{A}$ except the first one, γ is a hyperparameter that regularizes the degree of activation adjustment, and $\text{mean}(\mathbf{A}')$ and $\text{std}(\mathbf{A}')$ denote the mean and standard deviation of $\mathbf{A}'$. The resulting value, m^k, serves as the weight for the suppression and enhancement mask corresponding to the k-th token. We then apply the mask through element-wise multiplication with the tokens, yielding the activated tokens as $\mathbf{Z}' = \mathbf{Z} \odot \mathbf{M}$, where $\mathbf{M} = \{m^1, m^2, \cdots, m^N\}$.

Multi-Scale Token Selection

Considering that noisy tokens corresponding to complex background regions may misdirect the model's attention, the Multi-scale Token Selection (MSTS) mechanism is proposed to filter out noisy tokens. Let $\mathbf{Z'}_i = \{z_i^0, z_i^1, \cdots, z_i^{h_i \cdot w_i}\}$ denote the output from stage i. First, the `cls` token z_i^0 is detached, and the remaining patch tokens $\mathbf{P} = \{z_i^1, z_i^2, \cdots, z_i^{h_i \cdot w_i}\}$ are reshaped from $\mathbb{R}^{h_i \cdot w_i \times c_i}$ to $\mathbb{R}^{h_i \times w_i \times c_i}$. Next, we perform patch merging on $\mathbf{P}$ and flatten it to a 1D token sequence $\mathbf{P'} \in \mathbb{R}^{N_i \times c_i}$, where $N_i = (\frac{h_i \cdot w_i}{4})$. The patch merging method, inspired by Swin Transformer (Liu et al. 2021), expands the model's receptive field.

MSTS calculates scores $\mathbf{S}_i$ for tokens in $\mathbf{P'}$ to determine their importance and retains the top k_i tokens. The importance of each token $z_i \in \mathbb{R}^{c_i}$ is determined by averaging its activation across channels. The score s for each token is computed as:

$$s = \frac{1}{c_i} \sum_{j=1}^{c_i} z(j), \tag{5}$$

where c_i is the number of channels. The scores $\mathbf{S}_i = \{s^1, s^2, \cdots, s^{N_i}\}$ are sorted in descending order, and the top k_i indices $\mathbf{I}_i$ are selected. The corresponding tokens are then gathered from $\mathbf{P'}$ as follows:

$$\mathbf{I}_i = \text{topkIndex}(\mathbf{S}_i; k_i), \tag{6}$$

$$\hat{\mathbf{P}}_i = \text{gather}(\mathbf{P'}, \mathbf{I}_i). \tag{7}$$

To leverage the rich semantic information from deeper layers for localization, MSTS selects tokens in shallow layers correspond to those chosen in deep layers, along with layer-specific tokens. For MS-ViT, Stages 1 and 2 are designated as shallow layers, while Stage 3 serves as the deep layer. Stage 3 captures rich semantic information with its higher resolution map, enabling flexible token selection. For the `cls` token, instead of using the original `cls` token from each stage, we employ the `cls` Token Transfer (CTT) method (Moon et al. 2023) to leverage the rich semantic information from deeper layers. Specifically, a linear layer projects the `cls` token z_4^0 from the last stage to the `cls` tokens of preceding stages, as defined:

$$z_i^0 = \mathbf{W}_i^1 \left(\text{ReLU}\left(\text{BN}\left(\mathbf{W}_i^0 z_4^0\right)\right)\right). \tag{8}$$

Finally, the z_i^0 and $\hat{\mathbf{P}}_i$ are concatenated and passed through a Squeeze and Excitation network (Hu, Shen, and Sun 2018) to explore cross-channel interactions. They are then processed through standard Transformer Blocks independently to obtain multi-scale `cls` tokens $\hat{z}_i^0$ for classification:

$$\hat{\mathbf{Z}}_i = \text{Block}\left(\text{SEM}\left(\text{Concat}\left(z_i^0, \hat{\mathbf{P}}_i\right)\right)\right), \tag{9}$$

$$\eta_i = \mathbf{W}_i(\hat{z}_i^0), \tag{10}$$

where $\hat{z}_i^0$ is the `cls` token in $\hat{\mathbf{Z}}_i$.

Multi-Scale Dynamic Aggregation

One of the primary challenges in FGBR arises from the scale differences caused by close-up and long-distance shots, making it difficult to rely on a single scale for effective recognition. To address this, our method derives recognition results from multi-scale cues and aggregates them. However, simple summation aggregation indiscriminately combines results from all scales, potentially incorporating inappropriate or even harmful cues. To resolve this, we introduce a gating mechanism that selectively aggregates the most relevant results from multi-scale cues. Details of the mechanism are as follows:

First, we concatenate all `cls` token $\hat{z}_i^0$ to construct the global multi-scale features $\mathcal{MF}$ of the image. Simultaneously, we stack the recognition results from the four stages to obtain the global classification result $\mathcal{MC} \in \mathbb{R}^{4 \times n}$, where n is the number of the classes:

$$\mathcal{MF} = \text{Concat}(\hat{z}_1^0, \hat{z}_2^0, \hat{z}_3^0, \hat{z}_4^0), \tag{11}$$

$$\mathcal{MC} = \text{Stack}(\eta_1, \eta_2, \eta_3, \eta_4). \tag{12}$$

Next, we compute a set of gating weights $\mathbf{G} \in \mathbb{R}^{4 \cdot n}$ to guide the aggregation process:

$$\mathbf{G} = \sigma(\mathbf{W}\mathcal{MF} + b), \tag{13}$$

where $\mathbf{W}$ and b are learnable transformation weights and biases, and σ is the element-wise sigmoid activation function.

The gating weights $\mathbf{G}$ are then reshaped into $\mathcal{MG} \in \mathbb{R}^{4 \times n}$ to align with the multi-scale recognition results $\mathcal{MC}$. Finally, element-wise multiplication is performed, followed by a summation to produce the final recognition results $\mathcal{MR} \in \mathbb{R}^n$:

$$\mathcal{MR} = \sum_{i=1}^{4} \mathcal{MG}_i \cdot \mathcal{MC}_i. \tag{14}$$

The gating weights $\mathcal{MG}$ dynamically control the contribution of information from each scale, leveraging cues from

all scales. For a particular scale, the proportion of classification results included in the aggregation depends on the corresponding value in $\mathcal{MG}$. The larger the value, the more ideal the recognition result at that scale. This dynamic selection ensures that the model aggregates only the most appropriate classification results from different scales.

Additionally, bird species exhibit statistical differences in their scale distributions. Our gating mechanism aggregates cues across all scales and finely adjusts the weights of the classification scores on all classes from all scales, effectively addressing this challenge.

Loss Functions

With the output from the gating mechanism, we obtain five recognition results, $Y = \{y_1, y_2, y_3, y_4, y_5\}$. For each result y_i, we use the cross-entropy loss function with label-smoothing (Szegedy et al. 2016) for optimization, defined as:

$$\mathcal{L}_s = \sum_{y \in Y} \sum_{t=1}^{n} -\hat{y}_\beta^t \log y^t, \tag{15}$$

$$\hat{y}_\beta^t = \begin{cases} \beta, & t = \hat{t}, \\ \frac{1-\beta}{n}, & t \neq \hat{t}, \end{cases} \tag{16}$$

where n is the number of classes, t denotes the index of the label element, $\hat{t}$ is the index of the ground-truth class, and $\beta \in [0, 1]$ is a smoothing factor. We set β to incrementally increase in equal intervals of 0.1, ranging from 0.6 to 1.

To enhance diversity among features and enable multi-scale results to better capture different variations, we incorporate an extra contrastive loss, specifically for $\hat{z_4^0}$ and $\hat{z_3^0}$:

$$\begin{aligned} \mathcal{L}_{con} = \frac{1}{B^2} \sum_{i=1}^{B} [& \sum_{j:y_i=y_j}^{B} (1 - \cos(e^i, e^j)) + \\ & \sum_{j:y_i \neq y_j}^{B} (\max(\cos(e^i, e^j), 0)], \end{aligned} \tag{17}$$

where B denotes the batch size, $\cos(.,.)$ represents the cosine similarity function, and e^i and e^j refer to $\hat{z_4^0}$ or $\hat{z_3^0}$ extracted from the i-th and j-th images, respectively.

The final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_s + \alpha \mathcal{L}_{con}, \tag{18}$$

where α is the weighting factor for $\mathcal{L}_{con}$.

Experiments

Datasets

In our experiment, we demonstrate the effectiveness of our method on three fine-grained recognition datasets: CUB-200-2011 (Welinder et al. 2010), NABirds (Horn et al. 2015), and the iNat2017 (Horn et al. 2018). CUB-200-2011 is a widely used FGVC dataset, consists of $11,788$ images and 200 bird species, dividing $5,994$ images for training and $5,794$ images for testing. NABirds is a larger dataset, consists of $48,562$ images and 555 classes, dividing $23,929$ images for training and $24,633$ images for testing. iNat2017 is a large-scale dataset for fine-grained species recognition, consisting of $859,000$ images from over $5,000$ different species of plants and animals, dividing $579,184$ images for training and $95,986$ images for testing.

Method	Backbone	Acc (%)
NTS-Net (Yang et al. 2018)	ResNet50	87.5
Cross-X (Luo et al. 2019)		87.7
DCL (Wang et al. 2020)		87.8
PMG (Du et al. 2020)		89.6
MCEN (Li, Wang, and Zhu 2021)		89.3
GaRD (Zhao et al. 2021)		89.6
CMN (Deng et al. 2022)		88.2
MA-ASN (Zhang, Huang, and Liu 2022)		89.5
GDSMP-Net (Ke et al. 2023)		89.9
TA-CFN (Guan et al. 2023)		90.5
ViT (Dosovitskiy et al. 2021)	ViT-Base	90.6
RAMS-Trans (Hu et al. 2021)		91.3
AF-Trans (Zhang et al. 2022)		91.6
DCAL (Zhu et al. 2022)		92.0
TransFG (He et al. 2022a)		91.7
SIM-Trans (Sun, He, and Peng 2022)		91.8
IELT (Xu et al. 2023)		91.8
MpT-Trans (Wang, Fu, and Ma 2023)		92.0
ACC-ViT (Zhang et al. 2024)		91.8
MP-FGVC (Jiang et al. 2024a)		91.8
TransIFC+ (Liu et al. 2023)	Swin-Base	91.0
HERBS (Chou, Kao, and Lin 2023)		92.3
M2Former (Moon et al. 2023)	MS-ViT-Base	92.4
MDCM (Ours)		**92.5**

Table 1: Comparison with different methods on CUB-200-2011.

Implementation Details

For all experiments, we use the standard MS-ViT-Base model pre-trained on ImageNet21K as the backbone. Images are first resized to 600×600 pixels and then uniformly cropped to 448×448 pixels. For the training set, we apply random cropping and random horizontal flipping as data augmentation techniques, while for the test set, center cropping is employed. Our training strategy follows prior works (He et al. 2022a), employing an SGD optimizer with a momentum of 0.9 and a weight decay of 0.0. The batch size is set to 16 across all datasets. The learning rate is fixed at 0.045, with the contrastive loss weight set to 0.001. The hyperparameter γ in Eq 4 and α in Eq 18 are set as 0.3 and 0.1, respectively. We use cosine annealing for learning rate decay during training. All experiments are performed in a CUDA 11.0 environment with PyTorch 1.7.1, utilizing two NVIDIA RTX 3090 GPUs.

Main Results

Results on CUB-200-2011 The main results on CUB-200-2011 are shown in Table 1. Our method achieves a top-1 accuracy of 92.5%, demonstrating a substantial improvement over CNN-based approaches. When compared to ViT-based methods such as TransFG (He et al. 2022a), DCAL (Zhu et al. 2022) , and M2Former (Moon et al. 2023), our method outperforms them by 0.8%, 0.5%, and 0.1%,

Method	Backbone	Acc(%)
Cross-X (Luo et al. 2019)	ResNet50	86.4
PAIRS (Guo and Farrell 2019)		87.9
GaRD (Zhao et al. 2021)		88.0
CMN (Deng et al. 2022)		87.8
GDSMP-Net (Ke et al. 2023)		89.0
PMG-V2 (Du et al. 2022)	ResNet101	88.4
MGE-CNN (Zhang et al. 2019)		88.6
ViT (Dosovitskiy et al. 2021)	ViT-Base	89.6
TransFG (He et al. 2022a)		90.8
IELT (Xu et al. 2023)		90.8
MpT-Trans (Wang, Fu, and Ma 2023)		91.3
MP-FGVC (Jiang et al. 2024a)		91.0
ACC-ViT (Zhang et al. 2024)		91.4
TransIFC+ (Liu et al. 2023)	Swin-Base	90.9
M2Former (Moon et al. 2023)	MS-ViT-Base	91.1
MDCM (Ours)		**92.0**

Table 2: Comparison with different methods on NABirds.

Method	Backbone	Acc(%)
TASN (Zheng et al. 2019)	ResNet50	68.2
SRGN (Wang et al. 2024)		73.6
SSN(Recasens et al. 2018)	ResNet101	65.2
IARG(Huang and Li 2020)		66.8
RAMS-Trans (Hu et al. 2021)	ViT-Base	68.5
AF-Trans (Zhang et al. 2022)		68.9
SIM-Trans (Sun, He, and Peng 2022)		69.9
TransFG (He et al. 2022a)		71.7
MFVT (Lv et al. 2022)		72.6
ACC-ViT (Zhang et al. 2024)		77.0
M2Former (Moon et al. 2023)	MS-ViT-Base	77.8
MDCM (Ours)		**79.8**

Table 3: Comparison with different methods on iNat2017.

respectively. Furthermore, it achieves a 1.5% improvement over TransIFC (Liu et al. 2023), highlighting the effectiveness of our proposed method.

Results on NABirds The main results on the NABirds dataset are shown in Table 2. Our method achieves a top-1 accuracy of 92.0%, representing an improvement of at least 0.6%. Specifically, compared to recent ViT-based state-of-the-art methods (Zhang et al. 2024; Jiang et al. 2024a), our approach delivers improvements of 0.6% and 1.0%, respectively. When compared to the method (Moon et al. 2023) using the same backbone, we observe a 0.9% improvement. Additionally, against another bird recognition-focused method, our approach achieves a 1.1% improvement. These results substantiate the effectiveness of our proposed method.

Results on iNat2017 The main results on the large-scale iNat2017 dataset are shown in Table 3. Our method achieves at least a 2.0% improvement over other approaches, even when some of these methods employ significantly larger backbones (Ryali et al. 2023; He et al. 2022b). This highlights the efficiency and effectiveness of our approach.

Mechanisms	Score
Addition Operation	113
Gating Operation	**173**

Table 4: Impact of different aggregation choices in MSDA.

	MSTS	MSDA	MSCA	Acc(%)
(a)				91.6
(b)	✓			91.9
(c)	✓	✓		92.3
(f)	✓	✓	✓	**92.5**

Table 5: Ablation study of our MDCM on CUB-200-2011.

Ablation Study

Effectiveness of Proposed Modules We investigate the impact of the proposed modules, with results shown in Table 5. The baseline MS-ViT achieves a top-1 accuracy of 91.6% on CUB-200-2011. Introducing multi-scale cues from multiple stages (*i.e.,* Table 5(b)) results in a 0.4% improvement, demonstrating the effectiveness of our token selection design. Notably, for Table 5(b), the top-1 accuracy corresponds to the best accuracy among the pre-aggregation results, whereas in experiments with MSDA, it refers to the output of MSDA. Next, integrating our gating mechanism to dynamically aggregate recognition results from diverse multi-scale cues increases accuracy to 92.3% (*i.e.,* Table 5(c)). This finding highlights a limitation of prior methods, which process images at different scales across stages and rely on final features for downstream tasks, often failing to meet FGBR's need for diverse features. By contrast, multi-scale cues learned at different stages provide sufficient diversity. However, correct classifications are dispersed across these multi-scale features. MSDA dynamically aggregates the correct results, yielding a significant performance boost. Finally, applying MSCA to adjust the activation of cues in each stage delivers an additional 0.2% improvement, achieving a top-1 accuracy of 92.5%.

Contributions of MSDA We conduct additional experiments to analyze the effectiveness of the gating mechanism. First, we describe the evaluation approach for the gating mechanism. For each picture, five recognition results are generated, resulting in 2^5 possible combinations of correct or incorrect results. A robust aggregation choice should aim to maintain high accuracy in the aggregated classification result, even when some pre-aggregation results are incorrect. To quantify this, we assign correction scores ranging from -4 to $+4$ across the 32 possible combinations. The correction score for each image is computed as:

$$\text{Score} = \begin{cases} \sum_{i=1}^{4} +(y_i \doteq y), & \text{if } y_5 = y \\ \sum_{i=1}^{4} -(y_i \oplus y), & \text{if } y_5 \neq y \end{cases}, \tag{19}$$

Figure 3: The visualization of the MSTS mechanism highlights the selected tokens marked with red rectangles.

where y is the ground truth label, $\oplus$ denotes the exclusive OR operation, and $\doteq$ denotes the exclusive NOR operation. In brief, when the aggregated result is correct, the score reflects the number of incorrect pre-aggregation results that were corrected, assigning a corresponding positive value. Conversely, a negative score is assigned when the aggregated result is incorrect. This evaluation approach isolates the impact of pre-aggregation results, ensuring that the accuracy improvement is attributable to the aggregation mechanism rather than better pre-aggregation results. According to Table 4, even a simply summation of recognition results from multi-scale cues learned at different stages yields positive scores, demonstrating that these cues are sufficiently diverse to address the requirements of FGBR. Moreover, our gating mechanism adaptively aggregates these results, effectively mitigating the negative impact of inappropriate cues and achieving higher accuracy.

Impact of Different Stage Cues We further analyze the impact of cues at each stage on recognition accuracy. Following COCO dataset(Lin et al. 2014), we categorize the images in CUB-200-2011 based on bounding box sizes into three groups: `Small`, `Medium`, and `Large`, according to quartiles. Detailed results are presented in Table 6. The first row shows the performance of the baseline. The second row shows a slight improvement in accuracy, likely due to token selection reducing the influence of noise. Incorporating cues from Stage 3 into the recognition results enhances accuracy for medium and large-scale objects, indicating that Stage 3 cues provide valuable features for objects at these scales. Additionally, cues from Stage 2 and Stage 1 improve recognition across all object sizes—small, medium, and large. This suggests that shallow-layer detail information not only aids in identifying small objects but also supplements the recognition of medium and large objects. Overall, capturing diverse multi-scale cues from different stages effectively addresses the challenges of FGBR.

Cues from different stages				Acc(%)			
Stage 4	Stage 3	Stage 2	Stage 1	`Large`	`Medium`	`Small`	Total
				92.2	91.9	90.4	91.6
✓				91.7	92.2	90.5	91.7
✓	✓			92.5	92.5	90.6	92.0
✓	✓	✓		**92.9**	92.7	91.2	92.4
✓	✓	✓	✓	**92.9**	**92.8**	**91.5**	**92.5**

Table 6: The impact of cues learned at different stages

Visulization

Figure 3 shows the tokens selected at each stage by MSTS, with selected tokens highlighted by red rectangles. The first column displays the original image, while the subsequent columns, from left to right, represent tokens selected from the fourth stage to the first stage. In deeper layers, MSTS focuses on primary features at larger scales, whereas in shallower layers, it captures subtle or edge features. For instance, in the third sample, the deeper layers highlight the prominent yellow parts of the wings, while the shallower layers complement this by capturing finer details such as the eyes, beak, and the white feathers beneath the beak. Furthermore, with the guidance provided by MSTS, shallower modules effectively model deeper patches at smaller scales, enabling the capture of subtle differences. Overall, MSTS dynamically selects specific tokens across different stages, capturing discriminative representations in both shallow and deep layers, and providing rich cues for final decision-making.

Conclusion

This paper proposes a novel framework for fine-grained bird recognition, termed Multi-scale Diverse Cues Modeling (MDCM). The proposed framework captures diverse cues at different scales across various stages of a multi-scale Vision Transformer using an "Activation-Selection-Aggregation" paradigm. Specifically, the multi-scale cue activation module ensures that the discriminative cues learned at different stages are mutually distinct. Concurrently, a multi-scale token selection module is designed to remove redundant noise and emphasize discriminative, scale-specific cues at each stage. Finally, the selected tokens from each stage are independently utilized for bird recognition, with the recognition results adaptively fused through a multi-scale dynamic aggregation mechanism to make final model decisions. Qualitative and quantitative experiments consistently demonstrate the superiority of our MDCM for the FGBR task.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant 62332010) and the Major Research Program of Jiangsu Province (Grant BG2024042).

References

Abnar, S.; and Zuidema, W. H. 2020. Quantifying Attention Flow in Transformers. In *ACL*, 4190–4197.

Chou, P.; Kao, Y.; and Lin, C. 2023. Fine-grained Visual Classification with High-temperature Refinement and Background Suppression. *CoRR*, abs/2303.06442.

Deng, W.; Marsh, J.; Gould, S.; and Zheng, L. 2022. Fine-Grained Classification via Categorical Memory Networks. *IEEE Trans. Image Process.*, 4186–4196.

Dong, N.; Yan, S.; Tang, H.; Tang, J.; and Zhang, L. 2024a. Multi-view Information Integration and Propagation for occluded person re-identification. *Inf. Fusion*, 102201.

Dong, N.; Zhang, L.; Yan, S.; Tang, H.; and Tang, J. 2024b. Erasing, Transforming, and Noising Defense Network for Occluded Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.*, 4458–4472.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.; and Guo, J. 2020. Fine-Grained Visual Classification via Progressive Multi-granularity Training of Jigsaw Patches. In *ECCV*, 153–168.

Du, R.; Xie, J.; Ma, Z.; Chang, D.; Song, Y.; and Guo, J. 2022. Progressive Learning of Category-Consistent Multi-Granularity Features for Fine-Grained Visual Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9521–9535.

Fang, Z.; Jiang, X.; Tang, H.; and Li, Z. 2024. Learning Contrastive Self-Distillation for Ultra-Fine-Grained Visual Categorization Targeting Limited Samples. *IEEE Trans. Circuits Syst. Video Technol.*, 7135–7148.

Ge, W.; Lin, X.; and Yu, Y. 2019. Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up. In *CVPR*, 3034–3043.

Guan, X.; Yang, Y.; Li, J.; Zhu, X.; Song, J.; and Shen, H. T. 2023. On the Imaginary Wings: Text-Assisted Complex-Valued Fusion Network for Fine-Grained Visual Classification. *IEEE Trans. Neural Networks Learn. Syst.*, 5112–5121.

Guo, P.; and Farrell, R. 2019. Aligned to the Object, Not to the Image: A Unified Pose-Aligned Representation for Fine-Grained Recognition. In *WACV*, 1876–1885.

He, J.; Chen, J.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; and Wang, C. 2022a. TransFG: A Transformer Architecture for Fine-Grained Recognition. In *AAAI*, 852–860.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022b. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 15979–15988.

Horn, G. V.; Aodha, O. M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. J. 2018. The INaturalist Species Classification and Detection Dataset. In *CVPR*, 8769–8778.

Horn, G. V.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. J. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 595–604.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*, 7132–7141.

Hu, Y.; Jin, X.; Zhang, Y.; Hong, H.; Zhang, J.; He, Y.; and Xue, H. 2021. RAMS-Trans: Recurrent Attention Multi-scale Transformer for Fine-grained Image Recognition. In *ACM Multimedia*, 4239–4248.

Huang, Z.; and Li, Y. 2020. Interpretable and Accurate Fine-grained Recognition via Region Grouping. In *CVPR*, 8659–8669.

Jiang, X.; Tang, H.; Gao, J.; Du, X.; He, S.; and Li, Z. 2024a. Delving into Multimodal Prompting for Fine-Grained Visual Classification. In *AAAI*, 2570–2578.

Jiang, X.; Tang, H.; and Li, Z. 2024. Global Meets Local: Dual Activation Hashing Network for Large-Scale Fine-Grained Image Retrieval. *IEEE Trans. Knowl. Data Eng.*, 6266–6279.

Jiang, X.; Tang, H.; Yan, R.; Tang, J.; and Li, Z. 2024b. DVF: Advancing Robust and Accurate Fine-Grained Image Retrieval with Retrieval Guidelines. In *ACM Multimedia*, 2379–2388.

Ke, X.; Cai, Y.; Chen, B.; Liu, H.; and Guo, W. 2023. Granularity-aware distillation and structure modeling region proposal network for fine-grained image classification. *Pattern Recognit.*, 109305.

Li, G.; Wang, Y.; and Zhu, F. 2021. Multi-branch Channel-wise Enhancement Network for Fine-grained Visual Recognition. In *ACM Multimedia*, 5273–5280.

Li, Y.; Wu, C.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *CVPR*, 4794–4804.

Lin, D.; Shen, X.; Lu, C.; and Jia, J. 2015. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 1666–1674.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, volume 8693, 740–755.

Liu, H.; Zhang, C.; Deng, Y.; Xie, B.; Liu, T.; and Li, Y.-F. 2023. TransIFC: Invariant Cues-aware Feature Concentration Learning for Efficient Fine-grained Bird Image Classification. *IEEE Transactions on Multimedia*, 1–14.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*, 9992–10002.

Luo, W.; Yang, X.; Mo, X.; Lu, Y.; Davis, L.; Li, J.; Yang, J.; and Lim, S. 2019. Cross-X Learning for Fine-Grained Visual Categorization. In *ICCV*, 8241–8250.

Lv, X.; Xia, H.; Li, N.; Li, X.; and Lan, R. 2022. Mfvt: Multilevel feature fusion vision transformer and ramix data augmentation for fine-grained visual categorization. *Electronics*, 3552.

Malhotra, R. 2022. Habitat loss pushing more bird species to near extinction. *Nature India*, 18.

Moon, J.; Lee, J.; Lee, Y.; and Park, S. 2023. M2Former: Multi-Scale Patch Selection for Fine-Grained Visual Recognition. *CoRR*, abs/2308.02161.

Recasens, A.; Kellnhofer, P.; Stent, S.; Matusik, W.; and Torralba, A. 2018. Learning to Zoom: A Saliency-Based Sampling Layer for Neural Networks. In *ECCV*, 52–67.

Ryali, C.; Hu, Y.; Bolya, D.; Wei, C.; Fan, H.; Huang, P.; Aggarwal, V.; Chowdhury, A.; Poursaeed, O.; Hoffman, J.; Malik, J.; Li, Y.; and Feichtenhofer, C. 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. In *ICML*, 29441–29454.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *ACL*, 2931–2951.

Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; and Zeng, H. 2023. Git: Graph interactive transformer for vehicle re-identification. *IEEE Trans. Image Process.*, 1039–1051.

Shen, Y.; Sun, X.; Wei, X.; Jiang, Q.; and Yang, J. 2022. SEMICON: A Learning-to-Hash Solution for Large-Scale Fine-Grained Image Retrieval. In *ECCV*, 531–548.

Song, J.; and Yang, R. 2021. Feature Boosting, Suppression, and Diversification for Fine-Grained Visual Classification. In *IJCNN*, 1–8.

Sun, H.; He, X.; and Peng, Y. 2022. SIM-Trans: Structure Information Modeling Transformer for Fine-grained Visual Categorization. In *ACM Multimedia*, 5853–5861.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2818–2826.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. BlockMix: Meta Regularization and Self-Calibrated Inference for Metric-Based Meta-Learning. In *ACM Multimedia*, 610–618.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3Net: Multi-view Encoding, Matching, and Fusion for Few-shot Fine-grained Action Recognition. In *ACM Multimedia*, 1719–1728.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognit.*, 108792.

Wang, C.; Fu, H.; and Ma, H. 2023. Multi-Part Token Transformer with Dual Contrastive Learning for Fine-grained Image Classification. In *ACM Multimedia*, 7648–7656.

Wang, J.; Yu, X.; and Gao, Y. 2021. Feature Fusion Vision Transformer for Fine-Grained Visual Categorization. In *BMVC*, 170.

Wang, S.; Wang, Z.; Li, H.; Chang, J.; Ouyang, W.; and Tian, Q. 2024. Accurate Fine-Grained Object Recognition with Structure-Driven Relation Graph Networks. *Int. J. Comput. Vis.*, 137–160.

Wang, W.; Cui, Y.; Li, G.; Jiang, C.; and Deng, S. 2020. A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. *Neural Comput. Appl.*, 14613–14622.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD birds 200.

Xu, Q.; Wang, J.; Jiang, B.; and Luo, B. 2023. Fine-Grained Visual Classification via Internal Ensemble Learning Transformer. *IEEE Trans. Multim.*, 9015–9028.

Yan, R.; Xie, L.; Shu, X.; Zhang, L.; and Tang, J. 2023. Progressive instance-aware feature learning for compositional action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10317–10330.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to Navigate for Fine-Grained Classification. In *ECCV*, 438–454.

Zha, Z.; Tang, H.; Sun, Y.; and Tang, J. 2023. Boosting Few-Shot Fine-Grained Recognition With Background Suppression and Foreground Alignment. *IEEE Trans. Circuits Syst. Video Technol.*, 3947–3961.

Zhang, H.; Xu, T.; Elhoseiny, M.; Huang, X.; Zhang, S.; Elgammal, A. M.; and Metaxas, D. N. 2016. SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Recognition. In *CVPR*, 1143–1152.

Zhang, L.; Huang, S.; and Liu, W. 2022. Enhancing Mixture-of-Experts by Leveraging Attention for Fine-Grained Recognition. *IEEE Trans. Multim.*, 4409–4421.

Zhang, L.; Huang, S.; Liu, W.; and Tao, D. 2019. Learning a Mixture of Granularity-Specific Experts for Fine-Grained Categorization. In *ICCV*, 8330–8339.

Zhang, N.; Donahue, J.; Girshick, R. B.; and Darrell, T. 2014. Part-Based R-CNNs for Fine-Grained Category Detection. In *ECCV*, 834–849.

Zhang, P.; Dai, X.; Yang, J.; Xiao, B.; Yuan, L.; Zhang, L.; and Gao, J. 2021. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. In *ICCV*, 2978–2988.

Zhang, Y.; Cao, J.; Zhang, L.; Liu, X.; Wang, Z.; Ling, F.; and Chen, W. 2022. A free lunch from ViT: adaptive attention multi-scale fusion Transformer for fine-grained visual recognition. In *ICASSP*, 3234–3238.

Zhang, Z.; Chen, Z.; Wang, Y.; Luo, X.; and Xu, X. 2024. A vision transformer for fine-grained classification by reducing noise and enhancing discriminative information. *Pattern Recognit.*, 145: 109979.

Zhao, Y.; Yan, K.; Huang, F.; and Li, J. 2021. Graph-Based High-Order Relation Discovery for Fine-Grained Recognition. In *CVPR*, 15079–15088.

Zheng, H.; Fu, J.; Zha, Z.; and Luo, J. 2019. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. In *CVPR*, 5012–5021.

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022. Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. In *CVPR*, 4682–4692.

Zhuang, P.; Wang, Y.; and Qiao, Y. 2020. Learning Attentive Pairwise Interaction for Fine-Grained Classification. In *AAAI*, 13130–13137.