Disentangled Hypergraph-Guided Mamba Scanning for Fine-Grained Visual Recognition

Zhongwei Xiong[1,2*], Hao Wang[1,2*], Xiaoyan Yu[3†], Lingling Li[4,5], Xuezhuan Zhao[4,5], Taisong Jin[1,2]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China

[2]School of Informatics, Xiamen University, China

[3]School of Computer Science and Technology, Beijing Institute of Technology, China

[4]Zhengzhou University of Aeronautics, China

[5]Henan Provincial University-Enterprise R&D Center for Artificial Intelligence Technology, China

{zwxiong, haoge666}@stu.xmu.edu.cn, xiaoyan.yu@bit.edu.cn, {lilingling, zhaoxuezhuan}@zua.edu.cn, jintaisong@xmu.edu.cn

Abstract

Fine-grained Visual Recognition (FGVR) aims to distinguish between categories with subtle inter-class differences and large intra-class variations. While Vision Transformers with attention mechanisms have been widely adopted for FGVR, they usually suffer from high computational complexity and entangled global representations. Recent advancements in state-space models, exemplified by Mamba, have showcased substantial potential in vision-related tasks due to their linear scalability and rich sequence modeling capacity. To this end, we propose DHMamba, a novel Mamba based FGVR method. The proposed method leverages hypergraph to guide selective scanning and strengthen Mamba's capability in modeling fine-grained semantics. Furthermore, a Disentangled Local Scanning (DLS) module is introduced to utilize hyperedges to allocate distinct informative patches into independent channels for mitigating the representational entanglement. Extensive experiments conducted on multiple FGVR benchmarks demonstrate that the proposed DHMamba outperforms the state-of-the-art methods, validating the efficacy of combining state-space modeling with hypergraph-based feature structuring.

Introduction

Fine-grained Visual Recognition (FGVR) is dedicated to distinguishing between categories or instances that exhibit only subtle visual differences (Wei et al. 2022). Such challenges are prevalent in real-world applications, spanning from species of birds within the same genus to models of cars in the same series. In contrast to conventional image classification, FGVR faces the additional challenge of large intra-class variation stemming from environmental conditions, pose differences, and state changes. The recognition models must not only precisely localize the most informative regions of an image and effectively suppress interference from irrelevant backgrounds, but also extract discriminative features that capture the defining subtleties of fine-grained categories.

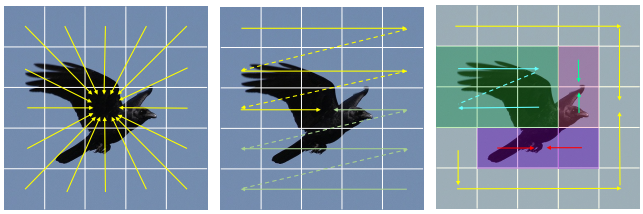
Early research (Qiu and Zhou 2020; Lin, RoyChowdhury, and Maji 2015; Yu et al. 2018) in FGVR was predominantly centered on Convolutional Neural Networks (CNNs) based architectures. These models sought to enhance the learning performance of subtle discriminative features through strategies such as part localization, feature refinement, and multi-branch learning (Lin, RoyChowdhury, and Maji 2015). With the advent of the attention mechanism, models that integrate attention modules with CNNs achieved significant progress in localizing crucial object parts and extracting salient features (Wu et al. 2019; Zheng et al. 2017; Sun et al. 2018), thereby facilitating more effective modeling of fine-grained categories. Nevertheless, the CNNs exhibit intrinsic limitations—most notably their dependence on deeply stacked hierarchical structures. Consequently, recent works (He et al. 2022; Hu et al. 2021; Zhang et al. 2022; Zhu et al. 2022; Xu et al. 2023; Chen et al. 2024) have shifted its focus toward Vision Transformers (ViTs). ViT (Dosovitskiy et al. 2021) partitions an image into a sequence of patches and models the relationships among them via a self-attention mechanism, enhancing the ability to capture subtle differences.

However, the self-attention mechanism incurs a computational complexity that is quadratic with respect to the number of patches, limiting the model's scalability to high-resolution images. Additionally, it tends to generate entangled feature representations, which hinders the identification and isolation of the most discriminative regions. To this end, some studies (Han et al. 2022; Chen et al. 2024) have begun to explore graph-based neural networks, serving either as feature encoding backbones or as plug-in modules, to enhance the representational capacity within the ViT framework. While the introduction of graph or hypergraph structures enhances the capability for modeling the complex data correlations, it can further exacerbate the problem of representation entanglement.

Recently, Mamba-based architectures (Gu and Dao 2024) have garnered considerable attention in computer vision (Zhu et al. 2024; Liu et al. 2024), owing to their lin-

*These authors contributed equally.

†Corresponding author.



(a) ViT (b) VMamba (c) Ours

Figure 1: The differences in patch processing among ViT-based models, VMamba, and our proposed DHMamba. ViT establishes attention-based interactions between each patch and all other patches, while VMamba adopts a bidirectional grid scanning strategy. In contrast, our proposed DHMamba employs a hypergraph-guided mechanism that enables Mamba to perform bidirectional scanning within each hyperedge, following the order of patch importance.

ear scalability and powerful sequence modeling capabilities. However, the in-depth application of Mamba-based models in FGVR remains under-explored based on the following two reasons: (1) The scanning strategy of Mamba typically employs a linear or bidirectional linear scanning mechanism, which tends to miss the critical semantic information in an image. (2) In scenarios where multiple spatially disjoint regions (e.g., a bird's head and claws) collectively provide discriminative cues for a class, Mamba tends to entangle these image regions, thereby limiting its capacity to learn robust feature representations.

To tackle the aforementioned chanllenges, we propose DHMamba, a novel mamba-based FGVR method. The proposed method leverages VMamba (Liu et al. 2024) as its image encoding backbone. Specifically, a hypergraph is proposed to guide the selective and informative scanning of VMamba for the FGVR task.

On the one hand, we propose a Local Information Gradient Scanning strategy to support a selective scanning of VMamba. This begins by constructing a hypergraph on the feature patches using Fuzzy C-Means clustering, followed by hypergraph convolution to model high-order semantic relations in an image. Furthermore, the node embeddings and the fuzzy membership matrix from clustering are leveraged to compute a semantic-aware score for each node. This score dictates a novel scanning path for the VSSBlock of Vmamba: the scan initiates at the node with the highest density and recursively progresses to the spatially adjacent node with the next-highest score.

On the other hand, we propose a Disentangled Local Scanning (DLS) module. This module leverages the semantic-aware score to assign hyperedges to different channels and guide Mamba to separately process distinct groups of patches. This approach facilitates the independent modeling of those regions containing core semantics (see Figure 1). To ensure effective disentanglement, we introduce a Maximum Orthogonality Loss, which explicitly encourages the sub-features generated from each independent scan to be semantically distinct and non-overlapping. Finally, these independently processed sequences are fused to form a comprehensive feature map for class prediction.

Our main contributions are summarized as follows:

- We propose DHMamba, a novel framework that integrates hypergraphs with Mamba for FGVR. To the best of our knowledge, our work is the first to explore the use of the Mamba architecture for pure FGVR.
- To enhance Mamba's ability to capture the fine-grained details, we propose a Local Information Gradient Scanning strategy. This strategy leverages hypergraph structures and semantic awareness to guide the patch scanning process along the most informative paths for feature extraction.
- To effectively model diverse and class-related semantic components within an image, we propose a Disentangled Local Scanning Module, which enables Mamba to independently process distinct sets of patches corresponding to different semantic sub-structures.

Related Work

Fine-grained Visual Recognition

CNNs have inspired numerous early approaches (Lin, RoyChowdhury, and Maji 2015; Zheng et al. 2017; Song and Yang 2021; Ding et al. 2021) for FGVR. For instance, PB R-CNN (Zhang et al. 2014) integrates R-CNN (Girshick et al. 2014) with spatial constraints to localize image regions and leverages high-resolution part features for improved classification. B-CNN (Branson et al. 2014) further employs a two-stream network to extract object locations and corresponding local features. Recently, with the success of self-attention mechanisms, ViT-based methods (Hu et al. 2021; Zhang et al. 2022; Zhu et al. 2022; Xu et al. 2023) have been explored to boost FGVR performance. For instance, TransFG (He et al. 2022) designs a Part Selection Module (PSM) that integrates all raw attention weights into a unified map to guide the selection of discriminative regions. HI2R (Chen et al. 2024) further incorporates a hypergraph-guided Structure Learning (HSL) module to capture high-order relations and address significant intra-category variation. However, the quadratic computational cost of self-attention in ViT remains a notable limitation for large-scale FGVR tasks. ViMD (Chen et al. 2025) explores the use of ViM for knowledge distillation in FGVR, but it does not address pure classification tasks. To this end, we propose DHMamba, which innovatively employs VMamba (Liu et al. 2024) to significantly reduce resource consumption while maintaining competitive classification performance.

Hypergraph Neural Networks

Hypergraph Neural Networks (HGNNs) have emerged as an effective extension of Graph Neural Networks (GNNs) for modeling high-order relationships. Unlike conventional pairwise graphs, hypergraph connects multiple nodes via hyperedges, naturally capturing multi-way interactions. HGNN (Feng et al. 2019) first generalizes convolution operations to hypergraph learning by applying hypergraph Laplacians and truncated Chebyshev polynomials in the

spectral domain. Subsequently, DHGNN (Jiang et al. 2019) improves representation quality by dynamically constructing task-specific hypergraphs. To overcome the limitations of spectral analysis in traditional matrix-based hypergraph representations, T-HyperGNN (Wang et al. 2024a) proposes a tensor-based framework that leverages hypergraph signal processing techniques. For vision tasks, Vision HGNN (Han et al. 2023), inspired by ViG (Han et al. 2022), partitions input images into patches as nodes to construct hypergraphs, successfully applying hypergraph convolution to general vision tasks with promising results. In the context of fine-grained visual recognition, hypergraph learning has shown strong potential to capture subtle relationships among parts or instances, enabling more robust discrimination under significant intra-class variation.

Methodology

Mamba Preliminaries

Selective Scan Structured State Space Sequence (S6) model (Gu and Dao 2024) is a category of sequence models that demonstrate strong capabilities to process sequential data. These models are primarily extensions of the previously proposed S4 model (Gu, Goel, and Re 2022), which projects an input stimulus $x(t) \in \mathbb{R}$ to an output response $y(t) \in \mathbb{R}$, through a hidden state $h(t)$. For continuous inputs, the system can be formulated using a set of linear ordinary differential equations:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t) + \boldsymbol{D}x(t), \end{aligned} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1\times N}$, and $\boldsymbol{D} \in \mathbb{R}^1$ are the weighting parameters.

By discretizing this system of ordinary differential equations, the continuous-time state space model can be adapted to handle discrete inputs. This is typically achieved using the zero-order hold (ZOH) discretization rule:

$$\begin{aligned} h_t &= \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t, \\ y_t &= \boldsymbol{C}h_t, \end{aligned} \tag{2}$$

where $\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A})$, $\overline{\boldsymbol{B}} = (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}$ are the transformation parameters used for discrete inputs and Δ denotes a learnable discretization step size. However, the linear time-invariant property of structured state space models imposes inherent limitations on their ability to adaptively capture contextual information. To overcome this constraint, Mamba (Gu and Dao 2024) extends the S4 model by introducing input-dependent parameters $\boldsymbol{B} = \boldsymbol{S}_B(x)$, $\boldsymbol{C} = \boldsymbol{S}_C(x)$, and $\Delta = \boldsymbol{S}_\Delta(x)$, thereby enabling a selective scanning mechanism to process the entire input sequence.

Recent studies (Zhu et al. 2024; Liu et al. 2024) have extended Mamba to vision tasks. VMamba (Liu et al. 2024) proposes a Cross Scan Module (CSM) that incorporates a 2D Selective Scan (SS2D) mechanism, which systematically scans the entire image from four different directions.

Overview of The Proposed Method

Figure 2 illustrates the overall architecture of our proposed model. Following VMamba's work, an input image first undergoes a patch-embedding layer, which transforms the image into $H \times W$ tokens. Subsequently, these tokens are fed into L Visual State Space Model (VSSM) layers. Each VSSM layer consists of serveal VSSBlocks and a downsampling layer. For each feature map before downsampling, it is input into the i-th layer of the proposed DHMamba module, which are also interconnected, forming cross-branch skip connections.

The DHMamba module consists of serveal Hypergraph-based Semantic Awareness (HSA) modules and a Disentangled Local Scanning (DLS) module. The HSA builds a hypergraph and outputs semantic-aware scores. To achieve the disentangling of hyperedges, K hyperedge clusters are formed by DLS, along with corresponding Disentangled VSSBlocks (DVSSBlocks). Subsequently, a downsampling operation is performed in synchronization with the VSSM. Finally, the two feature maps are concatenated and fed into a classifier.

Hypergraph-based Semantic Awareness Module

In this section, we propose the Hypergraph-based Semantic Awareness Module (HSA), which learns the high-order semantic associations among multiple image patches by constructing a hypergraph. HSA obtains a semantic awareness score via hypergraph structure for the subsequent disentangled guidance of VMamba scanning blocks.

Hypergraph Construction. We employ hypergraph to establish high-order semantic relationships among multiple image patches. Unlike traditional graph, which can only form connections between two nodes, hypergraph is a higher-order structure capable of connecting multiple nodes simultaneously. This characteristic makes hypergraph particularly suited for modeling complex associations among multiple nodes.

Formally, in a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, hyperedges $e_j \in \mathcal{E}$ each connect a subset of vertices, defined as $e_j = \{v_i \mid v_i \in \mathcal{V} \text{ and } i \in I_j\}$, where I_j is the set of indices for vertices that are included in hyperedge e_j. The set I_j directly corresponds to the nonzero entries of the j-th column of the incidence matrix $\boldsymbol{H} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{E}|}$, where $\boldsymbol{H}_{ij} = 1$ if vertex v_i is included in the hyperedge e_j.

We use Fuzzy-C Means (FCM) to first build a fuzzy hypergraph. Given the embeddings matrix of initial patches $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, where N is the number of input patches and d is the dimension of the embedding, and the number of cluster centers k, FCM initializes the membership matrix $\boldsymbol{U} = [u_{ij}] \in [0,1]^{N\times k}$, and the final value is obtained by iterating. Then, we construct the explicit hypergraph $\boldsymbol{H} \in \mathbb{R}^{N\times d}$ by hardening $\boldsymbol{U}$. Specifically, for each node v_i, we assign it to the cluster with the highest membership score, i.e., $j = \arg\max_j u_{ij}$. In this way, each cluster center corresponds to a hyperedge e_j that connects all nodes assigned to it.

Semantic Awareness of Image Patches. As previously discussed, images often contain multiple regions where different patches encapsulate key semantic information. To fully exploit the hypergraph structure in capturing these semantics and to facilitate subsequent operations such as dis-

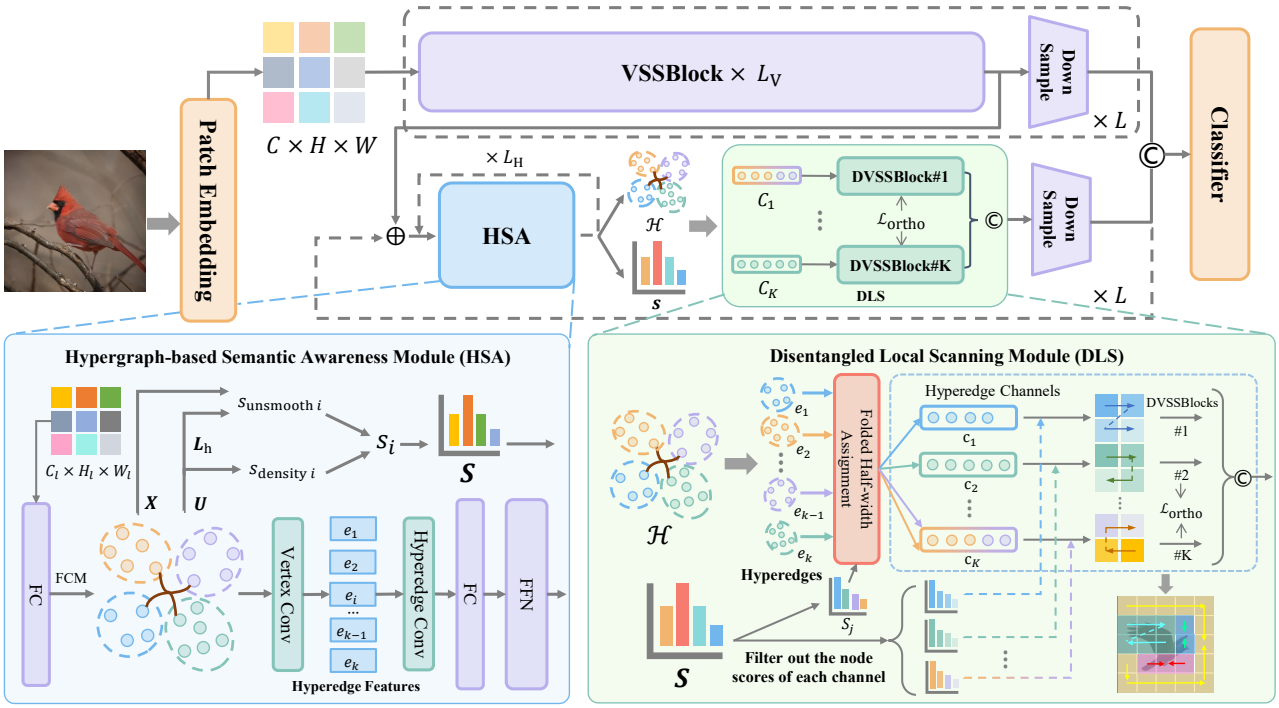


Figure 2: The overall architecture of our proposed model. The input image is first transformed into tokens of shape $C \times H \times W$, and then fed into the L layers, where C is the channel dimension. At the i-th layer, the feature map with shape $C_i \times H_i \times W_i$ modeled by L_{V} VSSBlocks is added to L_{H} HSA modules at the corresponding layer. After constructing the hypergraph $\mathcal{H}$, the feature matrix $\boldsymbol{X}$ and the membership matrix $\boldsymbol{U}$ are extracted. According to Eq. 3, Eq. 4 and Eq. 5, scores $\boldsymbol{S}$ are computed, and updated features are obtained through hypergraph convolution. Then, in the DLS module, scores S_j for the hyperedge e_j is derived based on $\boldsymbol{S}$, which guides the assignment of k hyperedges into K channels C_1 to C_K, each corresponding to a Disentangled VSSBlock. Herein, © denotes concatenation, and $\oplus$ indicates element-wise addition.

entanglement and scanning, we propose a semantic awareness scoring scheme based on information density. Intuitively, patches with high certainty (i.e., strongly associated with a specific cluster) are expected to carry more discriminative semantic information, while those with uncertain affiliations tend to be ambiguous or less informative.

Given the membership vector $\boldsymbol{u}_i = [u_{i1}, u_{i2}, ..., u_{ik}]$ for the i-th patch, its information density score s_{density_i} is defined as:

$$s_{\text{density}_i} = 1 - \frac{\mathcal{H}_i}{\log k}, \qquad \mathcal{H}_i = -\sum_{j=1}^{k} u_{ij} \log u_{ij}, \tag{3}$$

which reflects the semantic richness of the i-th patch within the context of its hyperedge.

Considering different image nodes in a hyperedge, if the node signal (i.e., its feature) corresponding to a certain image patch is smoother in the spatial neighborhood, it means that it is more likely to be a redundant patch. Leveraging the hypergraph Laplacian, we introduce the signal unsmoothness score s_{unsmooth_i} of the i-th patch as:

$$s_{\text{unsmooth}_i} = \text{Norm}(\|[\boldsymbol{L}_h \boldsymbol{X}]_i\|_2^2), \qquad \boldsymbol{L}_h = \boldsymbol{I} - \boldsymbol{D}_v^{-1/2} \boldsymbol{U} \boldsymbol{D}_e^{-1} \boldsymbol{U}^\top \boldsymbol{D}_v^{-1/2}, \tag{4}$$

where $\text{Norm}(\cdot)$ denotes normalization, $\boldsymbol{D}_v$ and $\boldsymbol{D}_e$ are the diagonal degree matrices for nodes and hyperedges, respectively.

We then compute the final semantic awareness score s_i:

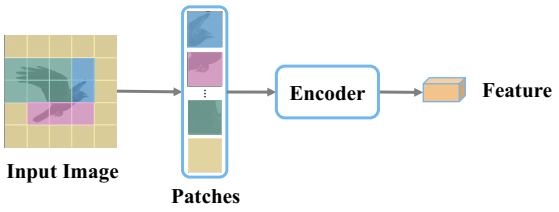
$$s_i = \lambda_{\text{s}} \cdot s_{\text{density}_i} + (1 - \lambda_{\text{s}}) \cdot s_{\text{unsmooth}_i}, \tag{5}$$

where $\lambda_{\text{s}} \in [0, 1]$ is a balancing hyperparameter. The semantic score $\boldsymbol{s} = [s_i] \in \mathbb{R}^N$ is forwarded to the next layer.

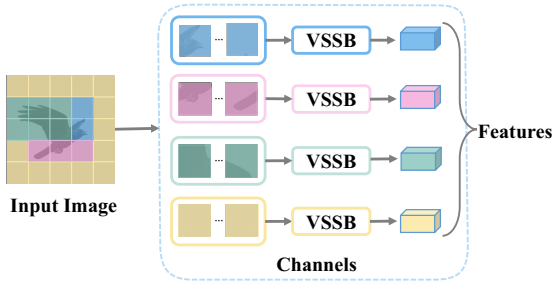
Hypergraph Convolution. To promote information exchange among multiple image patches connected by hyperedges, we adopt a classical two-step hypergraph convolution operation. Unlike standard graph convolutions that operate on pairwise edges, hypergraph convolution enables feature propagation over groups of semantically related nodes. Each hyperedge e_j aggregates features from its incident nodes first. And then, each node updates its representation by aggregating from the connected hyperedges:

$$\boldsymbol{X}' = \boldsymbol{D}_v^{-1/2} \boldsymbol{H} \, \sigma\left(\boldsymbol{W} \boldsymbol{D}_e^{-1} \boldsymbol{H}^\top \, \sigma\left(\boldsymbol{D}_v^{-1/2} \boldsymbol{X} \boldsymbol{\Theta}_1\right) \boldsymbol{\Theta}_2\right), \tag{6}$$

where $\boldsymbol{W} \in \mathbb{R}^{k \times k}$ is the diagonal matrix of hyperedge weights, $\boldsymbol{\Theta}_1 \in \mathbb{R}^{d \times N}$ and $\boldsymbol{\Theta}_2 \in \mathbb{R}^{N \times d}$ are learnable weight matrices, and $\sigma(\cdot)$ denotes a non-linear activation function. In between, $\boldsymbol{D}$ and $\boldsymbol{B}$ conduct normalization.



(a) Previous work using patch-based structures



(b) DHMamba (ours)

Figure 3: Compare to existing methods that typically use a unified encoder to learn a single representation from all image patches, our proposed DHMamba allocates patches into distinct channels to facilitate disentangled representation learning. Herein, VSSB refers to the Visual State Space Block used in VMamba.

Disentangled Local Scanning Module

The proposed Disentangled Local Scanning Module (DLS) aggregates image patches that belong to different important hyperedges into distinct hyperedge groups and performs VMamba scanning within each group (see Figure 3). Furthermore, scanning is conducted in the order of information density, allowing VMamba to preferentially absorb semantically rich content. In the following, we describes how the proposed method disentangles hyperedges into separate processing streams based on their semantic importance.

Hyperedges Disentanglement Since hyperedges naturally encode groups of patches sharing similar semantics, we aim to assign different hyperedges to separate channels for disentangled processing. However, to balance computational efficiency and semantic diversity, the number of available processing channels K is typically much smaller than the number of hyperedges k. Therefore, we propose a semantic-priority-based uniform allocation strategy, which preserves representation independence among informative patterns. Specifically, hyperedges carrying distinct cues are preferentially assigned to different output channels to ensure their disentangled representation. In contrast, less informative hyperedges can be compressed into shared channels.

Formally, we define the importance score S_j for the j-th hyperedge e_j as the average semantic awareness score of its constituent nodes:

$$S_j = \frac{1}{|e_j|} \sum_{i \in I_j} s_i, \tag{7}$$

where s_i is the score of node v_i (as defined in Eq. 5), and I_j is the index set of nodes in hyperedge e_j.

We sort the hyperedges in descending order of S_j to obtain a ranked list $e_{(1)}, e_{(2)}, ..., e_{(k)}$, where $e_{(1)}$ has the highest importance. Then, hyperedges are assigned to channel groups $\{C_1, C_2, \ldots, C_K\}$ in a round-trip fashion: starting from C_1 to C_K and then back to C_1, with the step length halved after each full forward-backward traversal. This process continues recursively until no further folding is possible. All remaining hyperedges are finally grouped into the last channel C_K.

To further encourage semantic disentanglement across channels, we introduce a Maximal Orthogonality Loss that promotes representational independence between hyperedge groups assigned to different channels. Specifically, for each pair of channels $c_1 \neq c_2$, we extract their aggregated feature representations $\boldsymbol{f}_{c_1}, \boldsymbol{f}_{c_2} \in \mathbb{R}^d$, and define the orthogonality loss as:

$$\mathcal{L}_{\text{ortho}} = \sum_{c_1 \neq c_2} \left| \frac{\boldsymbol{f}_{c_1}^\top \boldsymbol{f}_{c_2}}{|\boldsymbol{f}c_1| \cdot |\boldsymbol{f}c_2|} \right|^2. \tag{8}$$

The overall training objective becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{c}} + \lambda_{\text{ortho}} \cdot \mathcal{L}_{\text{ortho}}, \tag{9}$$

where $\mathcal{L}_{\text{c}}$ denotes the classification loss, and λ_{ortho} is a balancing weight.

Local Information Gradient Scanning Based on our hyperedge disentanglement above, the proposed Local Information Gradient Scanning strategy treats each hyperedge group independently and leverage semantic awareness scores to guide scanning within each group.

Let us consider one processing channel containing M hyperedges $e_1, e_2, ..., e_M$. Among them, we identify the most informative hyperedge $e_{\max} = \arg\max_{e_m}(S_m)$. Within $e_{\max}$, we locate the anchor node $v_0 = \arg\max_{v_i \in \mathcal{V}_{e_{\max}}}(s_i)$ as the node with the highest local score, where $\mathcal{V}_{e_{\max}}$ is the set of nodes in the hyperedge $e_{\max}$.

Starting from v_0, we construct a spatial scanning path by greedily selecting the most informative spatial neighbor at each step. Let $\mathcal{N}(v_t)$ denote the 4-connected grid neighbors of node v_t in an image. We define the scanning path $\mathcal{P} = [v_0, v_1, ..., v_T]$ such that:

$$v_{t+1} = \arg \max_{v \in \mathcal{N}(v_t) \cap \mathcal{V}_{\text{unvisited}}} s_v, \quad \text{subject to } s_v < sv_t, \tag{10}$$

where $\mathcal{V}_{\text{unvisited}}$ is the set of nodes not included in the path.

This recursive scanning process forms an information gradient path, along which semantic content flows from high-density to low-density regions. By feeding tokens in this ordered sequence into Mamba, we allow the model to progressively accumulate fine-grained visual patterns starting from the most semantically rich patch, enabling a structured and disentangled representation of visual concepts within each channel.

Experiments

Datasets

We evaluate our method on three widely used fine-grained image classification datasets:

Method	Venue	Backbone	CUB-200-2011	NABirds	Stanford Cars
PMG (Du et al. 2020)	ECCV 2020	ResNet	89.6	-	95.1
MCEN (Li, Wang, and Zhu 2021)	ACM MM 2021		89.3	-	95.2
GaRD (Zhao et al. 2021b)	CVPR 2021		89.6	88.0	95.1
CMN (Deng et al. 2022)	TIP 2022		88.2	87.8	94.9
ME-ASN (Zhang, Huang, and Liu 2021)	TMM 2022		89.5	-	94.8
SRGN (Wang et al. 2024b)	IJCV 2024		91.4	-	95.8
API-Net (Zhuang, Wang, and Qiao 2020)	AAAI 2020		88.6	86.6	94.9
PART (Zhao et al. 2021a)	AAAI 2020		90.1	-	95.3
CAL (Rao et al. 2021)	ICCV 2021		90.6	-	95.5
I2-HOFI (Sikdar et al. 2025)	IJCV 2025		90.1	91.0	94.3
CAP (Behera et al. 2021)	AAAI 2021	Xception	91.8	91.0	95.7
SR-GNN (Rao et al. 2021)	TIP 2022		91.9	91.2	96.1
ViT (Dosovitskiy et al. 2021)	ICLR 2020	ViT	90.6*	89.9*	93.5*
RAMS-Trans (Hu et al. 2021)	ACM MM 2021		91.3	-	-
AF-Trans (Zhang et al. 2022)	ICASSP 2022		91.5	-	95.0
TransFG (He et al. 2022)	AAAI 2022		91.7	90.8	94.8
DCAL (Zhu et al. 2022)	CVPR 2022		92.0	-	95.3
SIM-Trans (Sun, He, and Peng 2022)	ACM MM 2022		91.8	90.3*	-
IELT (Xu et al. 2023)	TMM 2023		91.8	90.8	-
MpT-Trans (Wang, Fu, and Ma 2023)	ACM MM 2023		92.0	-	93.8
MP-FGVC (Jiang et al. 2024)	AAAI 2024		91.8	91.0	-
HI2R (Wang, Fu, and Ma 2023)	ACM MM 2024		92.5	91.5	-
GKA (Wang et al. 2025)	IJON 2025		91.8	-	95.0
VMamba (Liu et al. 2024)	NeurIPS 2024	Mamba	89.7	88.7	94.5
DHMamba (Ours)	-	Mamba	**92.9**	**92.0**	**96.7**

Table 1: Comparison experiments with other state-of-theart methods on CUB-200-2011, NABirds and Stanford Cars datasets. All metrics are reported as Top-1 accuracy in percentage. Scores marked with * indicate reproduced results rather than those provided in the original papers. The best results are highlighted in bold.

CUB-200-2011 (Wah et al. 2011). The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset contains 11,788 images across 200 bird species. The images exhibit high intra-class variation and fine-grained differences in appearance, making it a challenging FGVR benchmark.

NABirds (Van Horn et al. 2015). The North American Birds (NABirds) dataset is a large-scale bird species dataset with 48,558 images from 555 categories. NABirds dataset has more subtle inter-class differences and a larger number of categories.

Stanford Cars (Krause et al. 2013). The Stanford Cars dataset consists of 16,185 images of 196 classes of cars, defined at the level of make, model, and year. Each class exhibits only subtle visual differences, such as variations in grille shape, headlight structure, or body lines.

All datasets are used in image classification settings, and we do not use any part annotations or bounding boxes.

Implementation Details

We adopt the `vmambav2-base-224 model` as the backbone for our framework, which is a pre-trained variant of VMamba trained on ImageNet-1K. This backbone uses a patch size of 4×4 and an initial embedding dimension of 128. The network architecture consists of four stages with VSSBlock layers in the configuration of $[2, 2, 9, 2]$, respectively.

To ensure fair comparisons with previous works, especially those based on ViT, we resize all input images to 600×600 and then apply a random crop to 448×448 during training. For testing, we use center cropping. Data augmentation includes widely adopted techniques such as random Gaussian blur, random grayscale conversion, random erasing, and color jittering.

The model is trained using the AdamW optimizer with a momentum of 0.9. We use a learning rate warm-up strategy for the first 20 epochs with an initial learning rate of 5×10^{-7}, which is linearly increased to the base learning rate of 5×10^{-4}. All models are trained for 300 epochs with a batch size of 16 across all datasets. We do not apply Exponential Moving Average (EMA) during testing.

For the HSA module, the number of clusters (i.e., hyperedges) k used during hypergraph construction is set to 20, and λ_{s} is set to 0.5. For the DLS module, the number of channels K is set to 10, and λ_{ortho} is set to 0.2. All experiments are implemented using PyTorch and conducted on 8 NVIDIA GeForce RTX 3090 GPUs.

Comparisons with State-of-the-art Methods

As mentioned above, we conduct the extensive comparisons with a range of state-of-the-art FGVR methods on three widely-used benchmarks: CUB-200-2011, NABirds and Stanford Cars. Top-1 accuracy is adopted as the perfor-

Ablation	CUB
Backbone	89.7
w/o Hypergraph convolution (HSA)	91.7
w/o Local information gradient scanning (DLS)	91.5
w/o Disentangled scanning (DLS)	91.0
DHMamba (Full)	92.9

Table 2: Effectiveness of different modules in DHMamba on CUB-200-2011 dataset, where without are abbreviated as w/o.

mance metric, and results are summarized in Table 1.

As shown in Table 1, the proposed DHMamba outperforms both CNN-based and Transformer-based counterparts, delivering superior performance across all datasets. On CUB-200-2011, DHMamba achieves an accuracy of 92.9%, outperforming the ViT-based methods such as HI2R (+0.4%) and DCAL (+0.9%). On NABirds, DHMamba reaches a new state-of-the-art accuracy of 92.0%, exceeding CAP (91.0%) and SR-GNN (91.2%) by a clear margin. On the Stanford Cars dataset, DHMamba achieves a leading level of 96.7%. These results validate the strong generalization ability and robustness of our model across diverse domains.

Notably, while VMamba slightly underperforms (-0.9% on CUB) compared to ViT as a linearly scaled encoder, our method still outperforms ViT-based methods by incorporating hypergraph-guided mechanism and disentangled scanning module, demonstrating the effectiveness of our module on Mamba. Furthermore, our method surpasses the compared methods that rely on sophisticated region localization (Yang et al. 2018; Zhao et al. 2021a) or part-based reasoning (Zheng et al. 2017; Du et al. 2020).

For computational efficiency, when image resolution is increased from 320×320 to 448×448, the training time multiplier for ViT is 1.77, whereas it is only 1.50 (-0.27) for our DHMamba based on the linear complexity of Mamba. This demonstrates that DHMamba can efficiently handle high-resolution FGVR tasks with lower computational overhead.

Ablation Experiments

Effectiveness of Key Components. We conduct ablation studies on the CUB-200-2011 dataset to evaluate the contributions of each major component in DHMamba, as summarized in Table 2.

Starting from the baseline Mamba backbone, which achieves 89.7% accuracy, we observe significant performance improvements after incorporating our proposed modules.

Removing the hypergraph convolution from the HSA module,i.e., allowing embeddings to pass directly to the next layer, resulting in a drop to 91.7%, confirming its role in capturing high-order semantic relations among patches.

Furthermore, omitting either the local information gradient scanning or the disentangled scanning (both from the DLS module) results in reduced accuracies of 91.5% and 91.0%, respectively. This demonstrates that both scanning enhancements are critical for effectively modeling spatial and semantic distinctiveness.

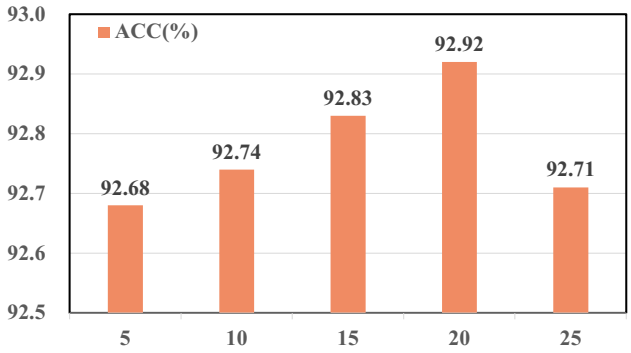


Figure 4: Top-1 accuracies (ACC) of different k in HSA module on CUB-200-2011 dataset.

Notably, we observe that the disentanglement contributes the most (+1.9%) to overall performance, as it significantly facilitates the identification of key patches.

When all components are combined in the full DHMamba model, we achieve the best performance, highlighting the complementary benefits of HSA and DLS in enhancing fine-grained visual recognition.

Influence of the Number of Clusters. As shown in Figure 4, we investigate how the number of clusters k in the HSA module affects performance on the CUB-200-2011 dataset. Since k controls the granularity of hyperedge construction in FCM clustering, it plays a crucial role in determining the quality of semantic-aware scanning.

The results show that increasing k from 5 initially improves accuracy by enabling finer semantic partitioning. However, when k becomes too large (exceeding 20), performance drops sharply, which may be attributed to excessive fragmentation of meaningful regions. The best performance (92.92%) is achieved when $k = 20$, indicating an optimal balance between discriminability and semantic coherence.

Conclusion

In this paper, we have proposed DHMamba, a novel Mamba-based framework tailored for FGVR. By addressing the limitations of existing methods—namely high computational complexity and entangled representations—we introduce Mamba to FGVR and propose two key innovations: (1) The Local Information Gradient Scanning strategy. It integrates hypergraph-based semantic awareness (HSA) into the scanning process, enabling more targeted and informative feature extraction. (2) The Disentangled Local Scanning (DLS) module. It allows the model to process semantically distinct regions in parallel, effectively capturing diverse, class-specific cues. Extensive experiments across multiple FGVR benchmarks show that the proposed DHMamba consistently outperforms state-of-the-art methods. In future research, we aim to achieve further advancements in both computational efficiency and classification accuracy metrics.

Acknowledgments

The work was supported by National Natural Science Foundation of China (No. 62072386), supported by Fujian Provincial Natural Science Foundation of China (No. 2025J01003), supported by Yunnan Provincial Major S&T Special Plan Project (No. 202402AD080001), supported by Henan Center for Outstanding Overseas Scientists (No. GZS2022011) and the Science and Technology Innovation Leading Talent Support Program of Henan Province, China(No.254200510017) , supported by Henan Province key research and development project (No. 231111212000) and Henan Province Higher Education Key Research Project (No. 25CY039) , supported by Open Found Project of Henan General Aviation Technology Key Laboratory (No.ZHKF-240205), supported by Chongqing Natural Science Foundation (No. CSTB2023NSCQ-MSX0070), supported by the Open Foundation of Key Lab of Oracle Information Processing of MOE (No. OIP2024E002). The authors gratefully acknowledge the technical support provided by Mr. Yanwen Liu, Engineer of the Information and Network Center of Xiamen University, in data processing for this study.

References

Behera, A.; Wharton, Z.; Hewage, P. R.; and Bera, A. 2021. Context-aware attentional pooling (cap) for fine-grained visual classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 929–937.

Branson, S.; Van Horn, G.; Belongie, S.; and Perona, P. 2014. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*.

Chen, L.; Wang, Q.; Li, Z.; and Yin, Y. 2024. Hypergraph-guided intra-and inter-category relation modeling for fine-grained visual recognition. In *Proceedings of the ACM International Conference on Multimedia*, 8043–8052.

Chen, Y.; Wang, J.; Wang, P.; Zhang, R.; and Li, Y. 2025. Vision mamba distillation for low-resolution fine-grained image classification. *IEEE Signal Processing Letters*, 32: 1965–1969.

Deng, W.; Marsh, J.; Gould, S.; and Zheng, L. 2022. Fine-grained classification via categorical memory networks. *IEEE Transactions on Image Processing*, 31: 4186–4196.

Ding, Y.; Ma, Z.; Wen, S.; Xie, J.; Chang, D.; Si, Z.; Wu, M.; and Ling, H. 2021. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, 30: 2826–2836.

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 *Proceedings of International Conference on Learning Representations*.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *Proceedings of European Conference on Computer Vision*, 153–168.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3558–3565.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 580–587.

Gu, A.; and Dao, T. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of Conference on Language Modeling*.

Gu, A.; Goel, K.; and Re, C. 2022. Efficiently modeling long sequences with structured state spaces. In *Proceedings of International Conference on Learning Representations*.

Han, K.; Wang, Y.; Guo, J.; Tang, Y.; and Wu, E. 2022. Vision gnn: An image is worth graph of nodes. *Advances in Neural Information Processing Systems*, 35: 8291–8303.

Han, Y.; Wang, P.; Kundu, S.; Ding, Y.; and Wang, Z. 2023. Vision hgnn: An image is more than a graph of nodes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19878–19888.

He, J.; Chen, J.-N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; and Wang, C. 2022. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 852–860.

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 *Proceedings of the ACM International Conference on Multimedia*, 4239–4248.

Jiang, J.; Wei, Y.; Feng, Y.; Cao, J.; and Gao, Y. 2019. Dynamic hypergraph neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2635–2641.

Jiang, X.; Tang, H.; Gao, J.; Du, X.; He, S.; and Li, Z. 2024. Delving into multimodal prompting for fine-grained visual classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2570–2578.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

Li, G.; Wang, Y.; and Zhu, F. 2021. Multi-branch channel-wise enhancement network for fine-grained visual recognition. In *Proceedings of the ACM International Conference on Multimedia*, 5273–5280.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 1449–1457.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in Neural Information Processing Systems*, 37: 103031–103063.

Qiu, C.; and Zhou, W. 2020. A survey of recent advances in cnn-based fine-grained visual categorization. In *Proceedings of IEEE International Conference on Communication Technology*, 1377–1384.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1025–1034.

Sikdar, A.; Liu, Y.; Kedarisetty, S.; Zhao, Y.; Ahmed, A.; and Behera, A. 2025. Interweaving insights: high-order feature interaction for fine-grained visual recognition. *International Journal of Computer Vision*, 133(4): 1755–1779.

Song, J.; and Yang, R. 2021. Feature boosting, suppression, and diversification for fine-grained visual classification. In *Proceedings of International Joint Conference on Neural Networks*, 1–8.

Sun, H.; He, X.; and Peng, Y. 2022. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the ACM International Conference on Multimedia*, 5853–5861.

Sun, M.; Yuan, Y.; Zhou, F.; and Ding, E. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision*.

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. 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/CVF Conference on Computer Vision and Pattern Recognition*, 595–604.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, C.; Fu, H.; and Ma, H. 2023. Multi-part token transformer with dual contrastive learning for fine-grained image classification. In *Proceedings of the ACM International Conference on Multimedia*, 7648–7656.

Wang, F.; Pena-Pena, K.; Qian, W.; and Arce, G. R. 2024a. T-HyperGNNs: Hypergraph neural networks via tensor representations. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3): 5044–5058.

Wang, S.; Wang, Z.; Li, H.; Chang, J.; Ouyang, W.; and Tian, Q. 2024b. Accurate fine-grained object recognition with structure-driven relation graph networks. *International Journal of Computer Vision*, 132(1): 137–160.

Wang, Y.; Ye, S.; Hou, W.; Xu, D.; and You, X. 2025. GKA: Graph-guided knowledge association for fine-grained visual categorization. *Neurocomputing*, 634: 129819.

Wei, X.-S.; Song, Y.-Z.; Aodha, O. M.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2022. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8927–8948.

Wu, L.; Wang, Y.; Li, X.; and Gao, J. 2019. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Transactions on Cybernetics*, 49(5): 1791–1802.

Xu, Q.; Wang, J.; Jiang, B.; and Luo, B. 2023. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 25: 9015–9028.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision*, 420–435.

Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; and You, X. 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European Conference on Computer Vision*.

Zhang, L.; Huang, S.; and Liu, W. 2021. Enhancing mixture-of-experts by leveraging attention for fine-grained recognition. *IEEE Transactions on Multimedia*, 24: 4409–4421.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-based R-CNNs for fine-grained category detection. In *Proceedings of the European Conference on Computer Vision*, 834–849.

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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 3234–3238.

Zhao, Y.; Li, J.; Chen, X.; and Tian, Y. 2021a. Part-guided relational transformers for fine-grained visual recognition. *IEEE Transactions on Image Processing*, 30: 9470–9481.

Zhao, Y.; Yan, K.; Huang, F.; and Li, J. 2021b. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15079–15088.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4692–4702.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the International Conference on Machine Learning*, 62429–62442.

Zhuang, P.; Wang, Y.; and Qiao, Y. 2020. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13130–13137.