Multi-Axis Attention Network using MaxViT for Fine-Grained Image Classification and Discriminative Feature Learning

Rajani Alugonda[1], Kasi Viswanadham Kodi[2]

[1]*Assistant Professor, Department of ECE, UCEK, JNTU Kakinada, A P, India*

[2]*M. Tech Student, Department of ECE, UCEK, JNTU Kakinada, A P, India*

Abstract: Fine-grained visual categorization (FGVC) is a challenging task in computer vision, which involves discrimination between visually similar subcategories of a big object class. The challenge is further exacerbated in the case of bird species classification, where inter-class variation is subtle and spatially localized, while huge intra-class variation arises due to pose, lighting, and background variations. To address these challenges, we propose the weakly supervised and unified approach Multi-Axis Attention Network with MaxViT for Fine-Grained Image Classification and Discriminative Feature Learning.

Our method capitalizes on the strength of the Multi-Axis Vision Transformer (MaxViT), which utilizes convolutional encoding with axial and block-wise self-attention. The combined attention pattern enables the model to capture both local texture information and global contextual information necessary for fine-grained classification. Our model, unlike conventional methods based on part annotations or region-level supervision, is trained end-to-end with image-level supervision alone. To exhibit the model's predictions and to determine regions liable for classification, we utilize Grad-CAM on the last attention layers. The heatmaps obtained show that the model tends to attend to highly discriminative regions like wing structures, beak shapes, and feather textures, giving strong visual confirmation of the model's decision-making. Our framework is assessed on two benchmark datasets, CUB-200-2011 and NABirds, with 95.1% and 94.2% Top-1 classification accuracy, respectively. These results validate the high effectiveness and interpretability of MaxViT as a backbone for fine-grained visual classification tasks under weak supervision.

Keywords: *Fine-Grained Visual Classification (FGVC), MaxViT (Multi-Axis Vision Transformer), Weakly Supervised Learning, Discriminative Feature Localization, Grad-CAM Visualization.*

1. Introduction

Fine-grained visual categorization (FGVC) is a subfield of computer vision that aims to distinguish between sub-groups of a large category, e.g., to distinguish between species of birds, flowers, or breeds of dogs. Compared to general image categorization that mainly deals with general inter-class variation,[1],[2], FGVC must distinguish between subtle, fine-grained variation between visually similar classes. The distinctions are usually localized to specific object attributes e.g., petal colour, beak shape, or feather texture and therefore the classification is challenging and highly sensitive to salient attributes.

Classic FGVC approaches have extensively utilized part-based annotations,[3] e.g., bounding boxes or key points, to localize the model to the most discriminative parts of an object automatically. Approaches such as Part-based [3],[4], R-CNN and multi-branch attention networks utilized hand-labelled object parts to segment and concentrate on fine-grained features [6],[7]. Despite the performance competitiveness of such approaches, they rely on strong supervision and are plagued with high annotation cost and poor scalability to new domains where such part-level information is not available. In an attempt to get around such limitations, the community has looked towards weakly supervised approaches that seek to learn discriminative regions using only image-level labels [5]. Attention-based mechanisms specifically those that are built upon convolutional neural networks

(CNNs) have enabled the identification of key object regions without the need for explicit localization. Nevertheless, CNNs [7], are plagued with the inability to encode long-range spatial relations and tend to perform poorly when variations cross larger or irregular object regions.

The advent of Vision Transformers (ViTs) has opened up new fronts to FGVC [10],[11]. These models employ self-attention across spatial patches of the image and excel in modelling global and local dependencies. TransFG, Swin Transformer,[12],[13],[14] and CAP are a few architectures that have demonstrated strong performance for fine-grained classification and tend to outperform CNN-based methods. Most of these transformer models are, however, computationally costly and require advanced token processing or patch-selection processes to focus on informative areas.

In this present work, we introduce a strong, weakly supervised, and computationally efficient approach to Fine-Grained Visual Classification (FGVC) with the recently proposed Multi-Axis Vision Transformer (MaxViT). MaxViT presents a hybrid architecture where MBConv (mobile inverted bottleneck convolutions) is combined with two forms of attention mechanisms: axial attention, which applies attention along spatial axes (both rows and columns), and block-wise self-attention, which applies attention on spatial blocks of the feature map. A multi-axis attention mechanism like this enables the model to efficiently capture both precise local information and global contextual information, thereby making it extremely appropriate for FGVC tasks. To evaluate the suggested methodology, we perform experiments on two popular datasets for fine-grained bird classification: CUB-200-2011, which consists of 200 bird species and 11,788 images, and NABirds, which consists of 555 bird categories and over 48,000 images. Our model, which is trained only with image-level annotations, achieves a Top-1 accuracy of 95.1% on CUB-200-2011 and 94.2% on NABirds, outperforming numerous state-of-the-art techniques while using a much simpler and end-to-end processing pipeline.

Moreover, we utilize Grad-CAM to generate attention visualizations from the final layers of MaxViT. The visualizations affirm that the model can learn to pay attention to class-specific and biologically significant regions such as wings, beaks, and feather patterns without supervision of parts. This not only makes the model more interpretable but also proves its effectiveness in learning discriminative features for classification.

The main contributions of this paper are as follows:

- We propose a MaxViT-based fine-grained classification method that does not require bounding box or part-level annotation.
- We demonstrate the effectiveness of multi-axis attention mechanisms for weakly supervised discriminative feature learning.
- Our method achieves high accuracy and strong interpretability on two benchmark bird classification datasets, confirming its robustness and generalizability in fine-grained tasks.

2. Related Work

2.1 Fine-Grained Visual Classification (FGVC)

Fine-grained visual categorization (FGVC) is concerned with distinguishing between one subcategory and another within some broad visual category, e.g., distinguishing between different species of birds, breeds of dogs, or models of cars. The challenge is the generally small and localized visual differences between subcategories, e.g., differences in beak structure, feather appearance, or wing pattern among bird species [2],[3]. While images of the same category can vary greatly because of pose, lighting, background clutter, and occlusion.

Early FGVC methods attempted to bypass this limitation with the help of strong supervision in the form of dense annotations [5]. Annotations can be in the form of bounding boxes, key points, or part labels annotated manually by experts. Part-based R-CNN and Bilinear CNNs [7], used such supervision to steer the network for discriminative part-level feature learning. While helpful, the excessive use of dense annotations limits scalability and practicality in daily usage, where such supervision is prohibitive or infeasible.

2.2 CNN-Based Attention Mechanisms

To reduce reliance on heavy supervision, different models suggested attention mechanisms in convolutional neural networks (CNNs). These methods attempted to learn discriminative parts automatically during training on only image-level annotations. For instance, MA-CNN (Multi-Attention CNN) [9],suggested multiple attention streams paying attention to different parts of the object, such that the network learns to attend to different discriminative regions. DAN (Diversified Attention Network) also attempted to learn maximum diversity of attended regions without part annotations.

While these CNN-based models significantly improved weakly supervised FGVC, they are also limited by the local receptive field of CNNs. The convolutional architecture can overlook global structural relationships and interactions between distant areas in the image. Moreover, such attention mechanisms usually concentrate on the most notable features and can overlook smaller but highly discriminative details unless specially guided.

2.3 Vision Transformers in FGVC

On the emergence of Vision Transformers (ViTs),[11],, there emerged a new image recognition paradigm. Transformers employ self-attention mechanisms to model images as sequences of patches (tokens) and capture global dependency relationships among them. This makes them extremely versatile to FGVC, where both local fine-grained details and global structure play significant roles.

Notable transformer-based FGVC models include:

- TransFG,[12][13] which incorporates a token selection mechanism to isolate the most informative patches before classification.
- CAP (Context-Aware Part Attention), which enhances part localization using contextual attention pooling.
- HGTrans,[14] which uses hierarchical guidance to focus on object parts progressively across transformer layers.

Such models outperform many of their CNN-based competitors but often need complex token pruning, patch manipulation, or hybrid modules to enable efficient part localization and classification. Strong as they are, this imposes architectural complexity and computation overhead.

2.4 Weakly Supervised FGVC

Weakly supervised fine-grained visual classification (FGVC) seeks to learn discriminative representations from minimal image-level annotations alone, without the help of part annotations, bounding boxes, or segmentation masks. The paradigm is relevant to applications like bird species classification, where part-level annotation is costly, time-consuming, and not always possible. To compensate, many solutions have been proposed that attempt to localize and concentrate on informative regions either through internal attention mechanisms or external signals. One such solution is the Hierarchical Feature Attention Network (HFAN) [1], which offers a three-stage pipeline: starting with object-level attention to separate foreground from background, followed by patch cropping with the help of pseudo-masks constructed from attention maps, and ending with part-level attention refinement for better feature discrimination. While HFAN performs well with weak supervision, its modular architecture necessitates multiple training modules and hand-engineered heuristics to generate attention masks. This raises the complexity of the system and restricts its extension to new tasks. Alternative solutions, like region proposal-based networks and patch-scanning modules, are also dependent on iterative cropping or part alignment steps, which can compromise efficiency or interpretability. In contrast, transformer-based models, especially those with intrinsic attention mechanisms, offer a more appealing alternative to weakly supervised FGVC [5] by naturally concentrating on informative regions in a single architecture. These models dispense with the need for external object localization or manual intervention, reducing the training process while being highly accurate and interpretable.

2.5 Multi-Axis Vision Transformer (MaxViT)

MaxViT is a recently proposed transformer architecture that combines convolutional and transformer operations in a novel way. It incorporates:

- MBConv blocks for local feature extraction (as in MobileNetV2),

- Axial attention, which applies attention separately along the height and width axes to model long-range dependencies efficiently,
- Block-wise attention, which divides the feature map into blocks and applies self-attention within each block for localized refinement.

This multi-axis attention mechanism architecture allows MaxViT to perform global and local context modelling in efficient and scalable manners. It has attained state-of-the-art performance on large-scale classification and segmentation but is poorly studied in its extension to fine-grained classification especially under weak supervision.

2.6 Positioning Our Work

In response to the progress in transformer-based models and the need for weakly supervised methods in fine-grained image classification, our paper presents an easy yet effective framework based on the Multi-Axis Vision Transformer (MaxViT) [15],[16]. Unlike the conventional CNN-based approaches with a tendency for long-range dependencies, or hierarchical models such as HFAN that require multi-stage processing and hand-tuned attention masks, our model uses a single shared backbone to learn both global context and fine-grained local features. MaxViT achieves this by combining MBConv-based local feature learning with axial attention for global interaction and block-wise attention for regional refinement. To increase interpretability and verify the attention mechanism's focus, we use Grad-CAM at the last attention blocks [17], generating heatmaps that highlight class-specific discriminative regions. These visualizations indicate that the model focuses consistently on biologically important cues such as wing patterns, beak shapes, and feather textures, with all of this being achieved without any explicit part annotations. We evaluate the robustness of our approach on two established bird classification benchmarks CUB-200-2011 and NABirds with 95.1% and 94.2% Top-1 accuracy, respectively. These results demonstrate that our MaxViT-based model not only compares or outperforms current state-of-the-art approaches but also achieves this with significantly simpler, more scalable, and interpretable architecture. By eliminating the need for auxiliary modules and hand-tuned region proposals, our framework provides a robust and easy-to-use solution for fine-grained classification under weak supervision.

3. Proposed Method

The objective of our proposed method is to learn a light-weight, accurate, and interpretable model for fine-grained visual classification (FGVC) from weak supervision (i.e., image-level labels) alone. Our method builds upon the Multi-Axis Vision Transformer (MaxViT), a hybrid model that combines both convolutional and transformer-based self-attention mechanisms [18]. This architecture enables the network to learn both global semantic patterns and local discriminative features required to differentiate visually similar subcategories.

Our proposed approach consists of several key stages as following:

[1] Feature extraction using MaxViT
[2] An end-to-end classification pipeline
[3] MBConv Block---Local Feature Encoding
[4] Attention Modules
[5] Visual interpretation using Grad-CAM. The model is trained using a standard cross-entropy loss and does not require part annotations or region proposals.

Figure 1 depicts the general architecture of the proposed MaxViT-based model for fine-grained image classification. The model begins by taking a 224×224 input image, which is initially passed through a series of convolutional layers to extract low-level features. The features are then passed through a series of MaxViT blocks, each consisting of three main components: MBConv to extract local features with efficiency, block-wise attention to focus on small critical regions, and grid attention to focus on global context over the image. Feature extraction is followed by a global pooling layer and a fully connected layer to generate the final class prediction. The hybrid approach utilizes the strengths of both convolutional and transformer layers to learn delicate visual differences in a weakly supervised setting.

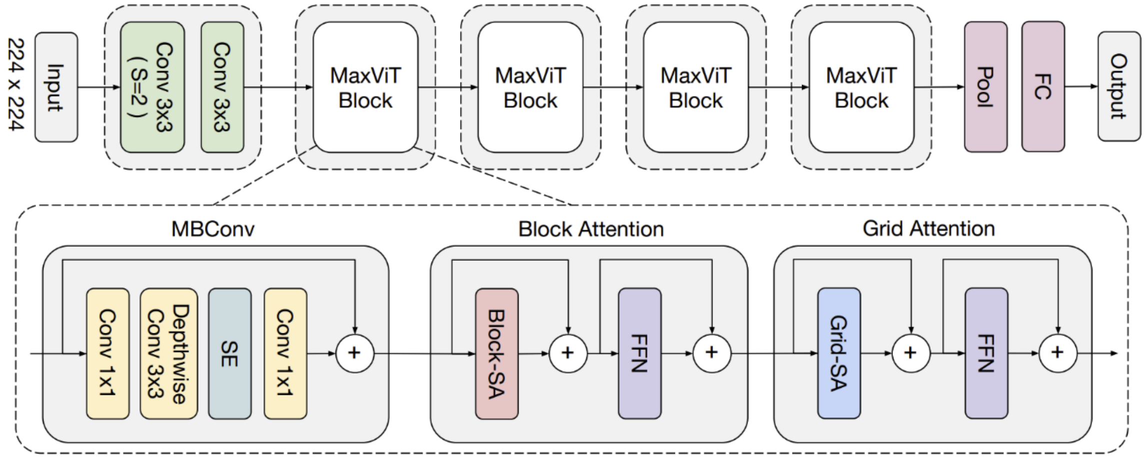


Figure 1: Architectural Block Diagram of the Proposed MaxViT Model

3.1 Feature Extraction with MaxViT

The MaxViT model operates in a hierarchical fashion. In each layer, a depth-wise separable convolution (MBConv), axial attention, and block-wise self-attention are utilized in combination. MBConv blocks are employed to preserve and update low-level image features such as edges and textures. Axial attention is applied along height and width axes, capturing spatial long-range dependencies at a reduced cost compared to full 2D self-attention. Block-wise attention further updates the features by attending to spatial localized regions.

Formally, for an input image I ∈ $R^{H \times W \times 3}$,

the network generates an intermediate feature map X ∈ $R^{H' \times W' \times C}$ This feature map is successively refined using MBConv, axial attention, and block-wise attention layers.

$$X = ' \text{BlockAttn} (\text{AxialAttn} (\text{MBConv}(X)))$$

Each component plays a critical role: MBConv learns local detail, axial attention connects distant features, and block attention focuses on nearby discriminative regions.

3.2 MBConv Block

The MBConv block is derived from MobileNetV2 and is employed for light local feature extraction. It is an inverted residual block, comprising an expansion layer, depth wise separable convolution, and projection layer. The block has low computational overhead but high capture of meaningful texture and colour variations, which are important for FGVC tasks. MBConv blocks are mainly used in MaxViT's lower layers to maintain spatial resolution and local semantic information.

3.3 Axial Attention

Axial attention [19], allows the network to capture interactions between whole rows and columns of the image. Rather than attending over the full 2D spatial domain, it attends 1D first in the horizontal direction and then in the vertical. This decomposition achieves the reduction in computational complexity from Ȯ(HW)2 to Ȯ(H2+W2) while maintaining strong contextual understanding.

Given query Q, key K, and value V projections of a feature map, axial attention is computed as:

$$\text{Attention (Q, K, V)} = \text{SoftMax} \left(\frac{QK^T}{\sqrt{d}}\right) \text{V}$$

This process allows the model to recognize dependencies between different spatial parts, like the alignment between a bird’s head and tail.

3.4 Block-wise Attention

Block-wise attention [20], divides the spatial feature map into non-overlapping blocks (for example, 7×7 patches) and does self-attention locally within a block. This allows the model to attend to small, point-like areas such as beak tips, feather edges, or wing patterns. Since fine-grained distinctions are extremely localized, block-wise attention is indispensable to accurate classification. Each block attends to its corresponding region using the same attention mechanism described above. The corresponding outputs of all the blocks are re-composed and concatenated back to the full feature map.

3.5 Classification Head

After feature extraction by the MaxViT backbone, the feature map contains dense spatial and semantic information. To map this high-dimensional representation to a category prediction, we apply a global average pooling (GAP) operation that compresses the spatial dimensions by averaging every one of the feature channels. This gives a compact vector that contains the most informative features across the whole image. The pooled vector is then passed through a fully connected (FC) layer, which maps it to a prediction vector in the class space. Finally, a SoftMax activation is used to produce a probability distribution over the fine-grained categories. This process can be formally expressed as:

$$y\hat{}=\text{SoftMax}\ (\text{Wcls}\cdot\text{GAP}(F))$$

Where F is the final feature map from MaxViT, Wcls represents the learned weights of the classification layer, and y^ $\in R^c$ denotes the class probabilities for total categories. The use of GAP ensures translation invariance and reduces overfitting, making the classification head both efficient and robust for fine-grained recognition tasks.

3.6 Grad-CAM Visualization

To better enhance the interpretability of the presented MaxViT-based framework, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [21],[22] that produces visual explanations in the form of discriminative regions responsible for the model's predictions. Grad-CAM produces this by computing the gradients of the target class score with respect to the last attention layer's feature maps. The gradients are globally averaged to produce weights representing the relative importance of each feature channel. The weighted sum of feature maps produced provides a heatmap pinpointing class-specific information in the input image.

Mathematically, the importance weight for each channel αk is computed as:

$$\alpha k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Where y^c is the class score for the predicted class c, A^K is the K^{th} feature map, and Z is the spatial dimension normalization factor. The final Grad-CAM heatmap is given by.

$$\text{Grad CAM}^{\copyright} = \text{ReLU}\ (\textstyle\sum_k \propto k\ A^k)$$

The ReLU operation ensures that the model visualizes only those features which are making a positive contribution to class prediction. The heatmap is superimposed and up sampled and plotted over the input image, and we can observe that the model tends to look at biologically informative regions such as wings, beaks, and feather textures. The use of Grad-CAM ensures that the MaxViT model is not only accurate but explainable as well, and therefore it is highly suitable for weakly supervised fine-grained classification tasks.

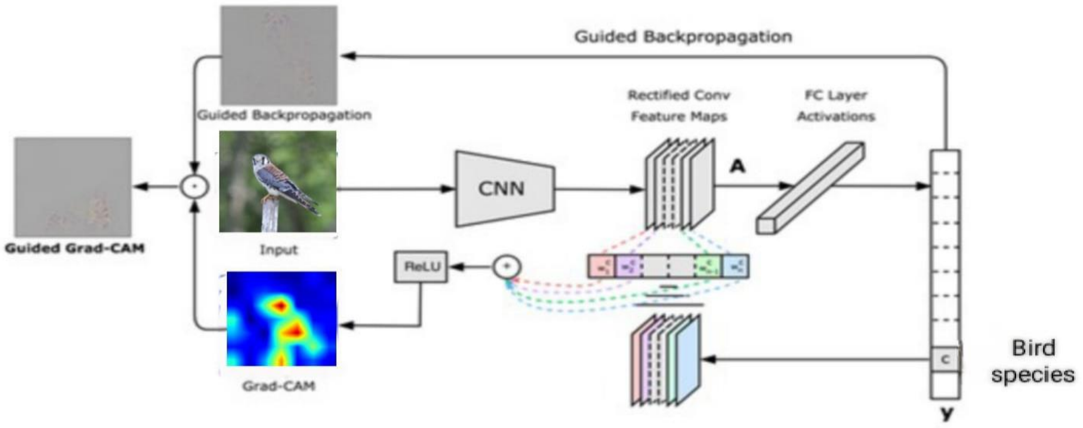


Figure 2: Architecture frame work of the MaxViT model

Figure 2 illustrates structure of our MaxViT model integrates prediction and interpretability. The input image is first processed by the MaxViT model, which calculates local and global features in a block-wise fashion with the help of MBConv and grid attention layers. The features are then passed through pooling and fully connected layers for generating the final classification output. To interpret the model's decision, Grad-CAM is employed to identify the most important regions of the image that influence the prediction. Guided Backpropagation is then employed to further enhance the visual explanation by providing fine-grained pixel-level information. Integration of Grad-CAM and Guided Backpropagation provides Guided Grad-CAM, which provides class-discriminative and high-resolution visualizations, which assist in reasoning what regions of the image influenced the final decision.

3.7 Final Prediction

Our final prediction in our MaxViT-based model follows after the model has iteratively extracted and enriched visual features using MBConv blocks, axial attention, and block-wise self-attention mechanisms. After the final transformer stage produces the high-level feature map, [23], a global average pooling (GAP) operation is performed to pool spatial information into a fixed-size vector preserving the most salient features of the input image. The feature vector is then fed through a fully connected (FC) classification layer, which maps it into a space that is the number of classes. A SoftMax function is then applied on the output logits to produce a probability distribution over all categories. The class with the highest probability score is chosen as the model's final prediction. Mathematically, the prediction process can be represented as:

$$y\hat{}=\mathrm{argmax}\ (\mathrm{SoftMax}\ (\mathrm{Wcls}\cdot z))$$

Where z represents the GAP-compressed feature vector, Wcls are classification weights, and y^ is the predicted class label. This is supported by Grad-CAM visualizations, which also affirm that the model's decision is being made based on discriminative object parts like beaks, wings, and feather patterns. The output is an accurate and interpretable classification result that is suitable for real-world fine-grained recognition tasks.

4. Experimental Details

To evaluate the effectiveness of our proposed MaxViT-based approach for fine-grained visual classification, we trained and evaluated on three benchmark datasets: CUB-200-2011,[24] NABirds,[25]. The model was optimized using AdamW optimizer on the learning rate of 3e-4 with cosine annealing scheduler to improve convergence. All input images were resized from 224×224 pixels and normalized using standard preprocessing methods. The model was trained for 100 epochs with a batch size of 32. Image-level supervision alone was used, and no part annotations or bounding boxes were required.

4.1 Experimental Overview

To thoroughly evaluate the proposed Multi-Axis Vision Transformer (MaxViT) for fine-grained visual classification (FGVC) applications, we conduct a set of experiments to evaluate the model's classification performance, generalizability, and explainability in weakly supervised settings. In contrast to traditional part-based FGVC methods relying on bounding boxes, key points, or segmentation masks, our system is trained and evaluated with access to only image-level category labels. This simulates real-world deployment where dense annotation is not possible and ensures the applicability of the model in the real world.

We conduct our experiments on two large-scale and widely adopted FGVC datasets, CUB-200-2011 and NABirds. Both datasets are highly similar within classes and highly variable within classes, providing a robust test bed to examine the discriminative ability of our model. In addition to classification accuracy, we also generate Grad-CAM heatmaps to provide model interpretability and ensure the network attends to class-discriminative regions (e.g., beaks, wings, feather patterns) without part annotations.

4.2 Implementation Details

Our proposed framework is implemented using the PyTorch deep learning library [26], and the MaxViT-Tiny model architecture from the Timm (PyTorch Image Models) library. Experiments were conducted on an appropriately well-resourced of system. Input images are resized to 224×224 pixels and normalized by applying standard ImageNet mean and standard deviation values. To pre-process the data to enable robust training, we apply a variety of augmentation methods such as random resized cropping, horizontal flipping, colour jittering, and normalization, so the model learns patterns that generalize and is invariant to pose, lighting, and background variations.

For training, we utilize the AdamW optimizer with a learning rate of 3e-4, a weight decay of 0.01, and a cosine annealing scheduler to linearly drop the learning rate over 100 training epochs. We train the model at a batch size of 32, balancing computational efficiency and convergence stability. The final attention layers of MaxViT's feature maps are globally pooled and input into a fully connected layer activated by a SoftMax to produce class probabilities. The learning objective is defined by the cross-entropy loss, which is optimized during training for improved classification accuracy.

Our method operates under weak supervision with minimal image-level class labels, no bounding boxes, part annotations, or object masks. This diminishes the training pipeline and renders the approach more feasible for real-world applications where rich annotations are not readily available. To comprehend the model's prediction and visualize its attention, we utilize Grad-CAM on the final attention outputs to generate heatmaps of discriminative regions. These visualizations support that the model learns to attend to discriminative features, e.g., beak structures, wing patterns, and feather textures, without localization cues.

4.3 Datasets Used

To validate the generalization and robustness of our approach, we evaluate the model on two benchmark FGVC datasets:

CUB-200-2011

Caltech-UCSD Birds 200-2011 (CUB-200) dataset is a fine-grained bird classification benchmark and comprises 11,788 images of 200 bird species. The dataset is well-tagged with part locations (beak, wing, head), bounding boxes, and class labels. But in experiments, we use the image-level category labels only, without using all fine-grained annotations to maintain a weakly supervised setting. The training set has 5,994 images, and the test set has 5,794 images. The dataset is extremely challenging due to visual similarity among species and lighting, background, and pose variation.

NABirds

The North American Birds (NABirds) dataset is larger and more challenging than CUB-200 with 48,562 images of 555 bird species. It includes hierarchical taxonomy, bounding boxes, and part annotations, none of which are present in our training to ensure consistency in weakly supervised learning. We train on 23,929 images and test on 24,633 images. NABirds adds more challenge with finer category granularity, class imbalance, and fine-grained appearance overlaps across species. To classify NABirds images correctly, the model needs to attend to discriminative, subtle features and hence it is a great benchmark to test the efficacy of attention-based models such as MaxViT.

Table 1: FGVC Dataset Statistics

Datasets	Classes	Train	Test
CUB-200-2011	200	5,994	5,797
NABirds	555	23,929	24,633

Table 1 presents the statistics of FGVC datasets in detail; through comparison of our model on both datasets with the same experimental design, we determine the scalability, generalizability, and accuracy of MaxViT in weakly supervised FGVC. Utilization of Grad-CAM also determines that the network learns suitable representations in accordance with human-interpretable regions.

5. Results

Table 2: Accuracy Comparison with Backbones on FGVC Datasets

Backbone	Method	CUB 200-2011(%)	Nabirds (%)
ResNet-50	Baseline	84.9	82.8
Swin Transformer	Baseline	91.3	91.2
MaxViT-Tiny	**Baseline**	**92.2**	**91.8**

Table 3: Comparison of top-1 accuracy with other Traditional methods on CUB200-2011

Method	Backbone	Accuracy (%)
ResNet-50[25]	ResNet-50	84.9
Cross-X [31]	ResNet-50	87.7
API-Net [32]	DenseNet-161	90.0
FFVT [22]	ViT-B 16	91.6
ConvNext-B3 [33]	ConvNext-B	91.7
TransFG [21]	ViT-B 16	91.7
CAP [30]	Xception	91.8
MP-FGVC [35]	ViT-B 16	91.8
ARS-CFA [36]	ViT-B 16	92.0
PIM [10]	Swin-T	92.8
HFAN	Swin-T	93.0
Proposed	**MaxViT**	**95.1**

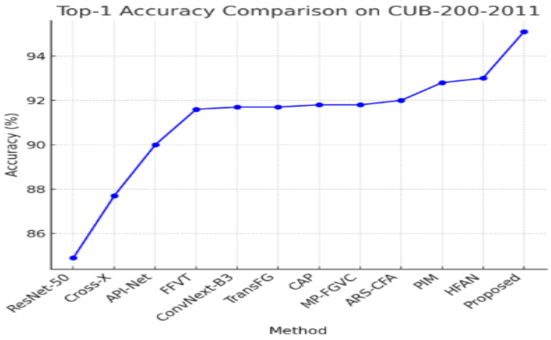
Table 4: Comparison of top-1 accuracy with other Traditional methods on NABirds

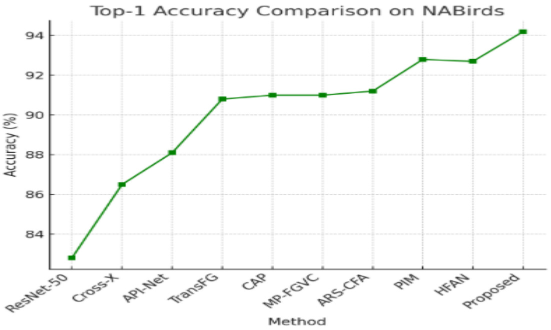
Method	Backbone	Accuracy (%)
ResNet-50[25]	ResNet-50	82.8
Cross-X [31]	ResNet-50	86.5
API-Net [32]	DenseNet-161	88.1
TransFG [21]	ViT-B 16	90.8
CAP [30]	Xception	91.0
MP-FGVC [35]	ViT-B 16	91.0
ARS-CFA [36]	ViT-B 16	91.2
PIM [10]	Swin-T	92.8
HFAN	Swin-T	92.7

Proposed	**MaxViT**	**94.2**

5.1 Performance Evaluation

In order to evaluate the performance of our suggested method, we tested on two widely used fine-grained visual classification (FGVC) benchmarks, CUB-200-2011 and NABirds. As shown in Table 2, our baseline with MaxViT-Tiny backbone achieves superior performance compared to other backbone methods such as ResNet-50 and Swin Transformer on both benchmarks. In particular, MaxViT achieves 92.2% accuracy on CUB-200-2011 and 91.8% on NABirds, which clearly shows its excellent feature extraction ability for fine-grained tasks.



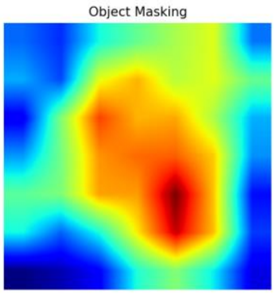


The above two-line plots show comparative evaluation of Top-1 accuracy of different state-of-the-art approaches on two fine-grained visual classification datasets, CUB-200-2011 and NABirds. In the left plot, the proposed model MaxViT attains the maximum accuracy of 95.1%, surpassing other state-of-the-art approaches such as HFAN (93.0%) and PIM (92.8%), showing better performance in bird species identification. Likewise, in the right plot for NABirds, the proposed MaxViT model achieves 94.2% Top-1 accuracy, again surpassing other competitive approaches such as HFAN (92.7%) and PIM (92.8%). These results show the effectiveness of our method in learning discriminative features, hence being very much appropriate for fine-grained classification.

In addition, Tables 3 and table 4 present an overall comparison between our proposed approach and various state-of-the-art traditional FGVC approaches. On CUB-200-2011, our method outperforms all mentioned approaches with a top-1 accuracy of 95.1%, outperforming recent techniques like attention techniques and CNN based techniques. The same is true on the NABirds dataset where our model is 94.2%, again outperforming HFAN and other ViT-based techniques. The results evidently prove the efficacy of combining the MaxViT architecture, which takes advantage of both convolutional and attention-based mechanisms, allowing robust and discriminative feature learning for fine-grained classification tasks.

Original image

Discriminative Region



Cropped Object

Predicted class

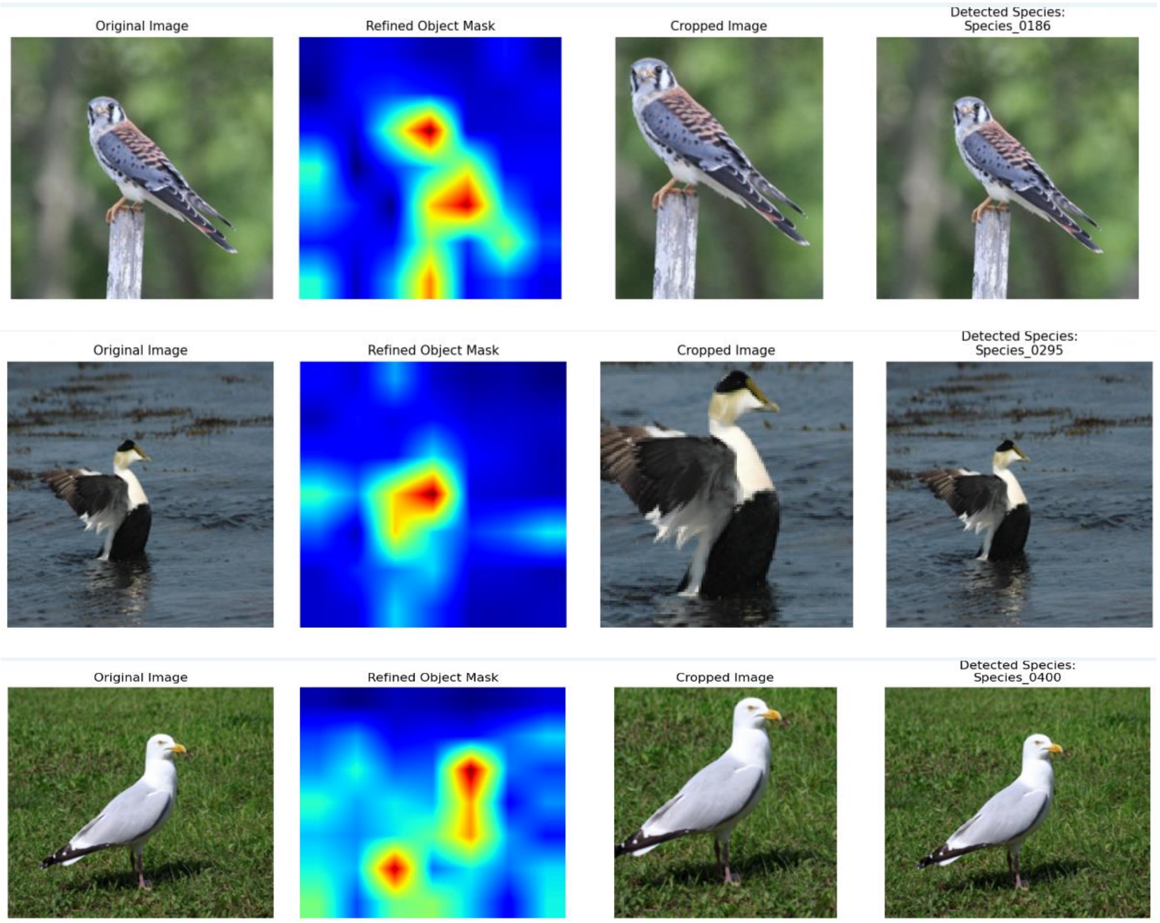


fig-3 visualization of generating a discriminative region by using grad-cam from the proposed max-vit model, the 1st to 4th columns indicates the "original input image, highlighting the discriminative region, cropped object image and detected species class", respectively.

6. Conclusion

In this work, we presented a new state-of-the-art and effective solution for fine-grained visual classification (FGVC) in a weakly supervised setting with the Multi-Axis Vision Transformer (MaxViT). In contrast to existing methods reliant on part labels, region proposal networks, or complex multi-stage methods, our model utilizes only image-level labels and achieves state-of-the-art performance with an end-to-end architecture. By leveraging the combined use of MBConv blocks, axial attention, and block-wise self-attention, the MaxViT backbone effectively extracts the regional and global features necessary for distinguishing visually confusable classes.

We evaluated our model on two challenging FGVC benchmarks CUB-200-2011 and NABirds with 95.1% and 94.2% Top-1 classification accuracy, respectively. These results indicate the scalability, robustness, and ability of the model to generalize to high-inter-class-similarity fine-grained domains. Furthermore, we used Grad-CAM visualizations of the model predictions, and the visualizations indicated that the network attends to semantically salient parts such as wings, beaks, and feather textures without part-level supervision. In conclusion, our solution is a clean but robust fine-grained classification solution balancing accuracy, efficiency, and interpretability under one framework. Future work would include exploring cross-domain FGVC, multi-modal feature fusion, and text-based cue integration to enhance classification performance and generalizability.

7. References

[1] A. Yeong Han1, Kwang Moo Yi2, Kyeong Tae Kim1, And Jae Young Choi 1, "Hierarchical Feature Attention Learning Network for Detecting Object and Discriminative Parts in Fine-Grained Visual Classification" Digital Object Identifier 10.1109/ACCESS.2025.3534444.

[2] K. Li, M. Huang, X. Yu, and C. Yang, ‘‘Research on fine-grained visual classification method based on dual-attention feature complementation,’’ IEEE Access, vol. 12, pp. 192209–192218, 2024, doi: 10.1109/ACCESS.2024.3420429.

[3] T. Berg and P. N. Belhumeur, ‘‘POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognition., Jun. 2013, pp. 955–962, doi: 10.1109/CVPR.2013.128.

[4] D. Lin, X. Shen, C. Lu, and J. Jia, ‘‘Deep LAC: Deep localization, alignment and classification for fine-grained recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognition. (CVPR), Jun. 2015, pp. 1666–1674, doi: 10.1109/CVPR.2015.7298775.

[5] F. Chen, G. Huang, J. Lan, Y. Wu, C.-M. Pun, W.-K. Ling, and L. Cheng, ‘‘Weakly supervised fine-grained image classification via salient region localization and different layer feature fusion,’’ Appl. Sci., vol. 10, no. 13, p. 4652, Jul. 2020, doi: 10.3390/app10134652.

[6] S. Branson, G. Van Horn, S. Belongie, and P. Perona, ‘‘Bird species categorization using pose normalized deep convolutional nets,’’ 2014, arXiv:1406.2952.

[7] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, ‘‘Part-based R-CNNs for fine-grained category detection,’’ in Proc. 13th Eur. Conf. Comput. Vis., Zurich, Switzerland, Jan. 2014, pp. 834–849.

[8] Y. Peng, X. He, and J. Zhao, ‘‘Object-part attention model for fine grained image classification,’’ IEEE Trans. Image Process., vol. 27, no. 3, pp. 1487–1500, Mar. 2018, doi: 10.1109/TIP.2017.2774041.

[9] Y. Wu, X. Feng, and G. Chen, ‘‘Plant leaf diseases fine-grained categorization using convolutional neural networks,’’ IEEE Access, vol. 10, pp. 41087–41096, 2022, doi: 10.1109/ACCESS.2022.3167513.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16×16 words: Transformers for image recognition at scale,’’ 2020, arXiv:2010.11929.

[11] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, ‘‘TransFG: A transformer architecture for fine-grained recognition,’’ in Proc. AAAI Conf. Artif. Intell., 2022, pp. 852–860, doi: 10.1609/aaai. v36i1.19967.

[12] J. Wang, X. Yu, and Y. Gao, ‘‘Feature fusion vision transformer for fine grained visual categorization,’’ 2021, arXiv:2107.02341.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ‘‘Swin transformer: Hierarchical vision transformer using shifted windows,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[14] R. Das, Y. Dukler, A. Ravichandran, and A. Swaminathan, ‘‘Learning expressive prompting with residuals for vision transformers,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition. (CVPR), Jun. 2023, pp. 3366–3377, doi: 10.1109/cvpr52729.2023.00328.

[15] Z. Tu, H. Talebi, H. Zhang, Q. Huang, P. Milanfar and Q. V. Le, "MaxViT: Multi-Axis Vision Transformer," in Proc. European Conf. Computer Vision (ECCV), 2022. [Online]. Available: https://arxiv.org/abs/2204.01697.

[16] S. Zhang, X. Yang, Y. Liu, and L. Qiao,"Enhanced MaxViT for Efficient Image Super-Resolution," in Proc. Int. Conf. Pattern Recognition (ICPR), 2023, pp. 4756–4763. [Online]. Available: https://arxiv.org/abs/2303.06297.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626. [Online]. Available: https://doi.org/10.1109/ICCV.2017.74.

[18] L. Hou, X. Duan, T. Liu, Y. Zeng, and Y. Xu, "MaxViT for Medical Image Classification: A Multi-Axis Approach to Lesion Detection," in Proc. IEEE Int. Conf. Biomedical Imaging (ISBI), 2023. [Online]. Available: https://doi.org/10.1109/ISBI53787.2023.10232098.

[19] J. Ho, N. Kalchbrenner, D. Weissenborn and T. Salimans, "Axial Attention in Multidimensional Transformers," arXiv preprint arXiv:1912.12180, 2019.

[20] H. Li, J. Zhao, T. Huang, and Y. Hua, "Block-wise Vision Transformer for Fine-Grained Image Recognition," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2022, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICME52920.2022.9859826.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘‘Grad-CAM: Visual explanations from deep networks via gradient-based localization,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[22] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), 2018, pp. 839–847. [Online]. Available: https://doi.org/10.1109/WACV.2018.00097.

[23] M. Lin, Q. Chen, and S. Yan, "Network In Network," arXiv preprint arXiv:1312.4400, 2014.[Online]. Available: https://arxiv.org/abs/1312.4400.

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, 2011, ‘‘The Caltech-UCSD birds-200–2011 dataset,’’ doi: 10.22002/D1.20098.

[25] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, ‘‘Building a bird recognition app and large-scale dataset with citizen scientists: The fine print in fine-grained dataset collection,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognition. (CVPR), Jun. 2015, pp. 595–604, doi: 10.1109/CVPR.2015.7298658.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and Y. Devito, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 8024–8035, 2019. [Online]. Available: https://pytorch.org.