

1 Feature Analysis and Interpretability for Fine-Grained Gull Classification

1.1 Methodology for Wing and Wingtip Intensity Analysis

To quantitatively validate the visual differences in wing and wingtip coloration between Slaty-backed Gulls and Glaucous-winged Gulls, we developed a comprehensive image analysis pipeline. The methodology below is a precise reflection of our implemented codebase and the sequence of analytical steps it performs.

1.1.1 Image Selection and Preparation

High-resolution images of both species were selected for analysis. Each image was accompanied by a manually created segmentation mask, generated in Adobe Photoshop, with the following color codes:

- **Red** (RGB: 255,0,0): Wing region
- **Green** (RGB: 0,255,0): Wingtip region

Only the wing and wingtip regions were analyzed in this pipeline. Images were assumed to be appropriately filtered for quality and relevance prior to analysis.

1.1.2 Region Extraction and Intensity Normalization

For each image:

1. Both the original image and its segmentation mask were loaded using OpenCV.
2. The original image was converted to grayscale and min-max normalized to the [0,255] range to account for variations in lighting conditions.
3. Pixels corresponding to the wing and wingtip regions were extracted from the normalized grayscale image using color thresholding on the segmentation mask.

1.1.3 Wing Region Intensity Analysis

For each extracted wing region, the following statistics were computed:

- Mean intensity
- Standard deviation
- Median, minimum, and maximum intensity values
- Skewness and kurtosis of the intensity distribution

- Total pixel count in the region

These metrics were saved for all images and subsequently aggregated by species for summary statistics and statistical testing.

1.1.4 Wingtip Region Intensity and Contrast Analysis

For each extracted wingtip region, the following analyses were performed:

- **Mean, standard deviation, median, min, and max intensity values**
- **Pixel intensity distribution:** The percentage of wingtip pixels falling into 10-unit intensity bins across the full range (0–10, 10–20, ..., 240–255)
- **Very dark pixel analysis:** The percentage and count of wingtip pixels below absolute intensity thresholds of 30, 40, 50, and 60 (`pct_dark_lt_30`, `pct_dark_lt_40`, etc.)
- **Wing–wingtip contrast:** For each wingtip pixel, the difference from the mean wing intensity was computed. The percentage of wingtip pixels darker than the mean wing intensity (`pct_darker_pixels`) was calculated, as well as the percentage of pixels exceeding difference thresholds of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 (`pct_diff_gt_X`).

This multi-metric approach quantitatively characterizes both the absolute darkness of the wingtip and its contrast relative to the rest of the wing.

1.1.5 Statistical Analysis and Visualization

- **Species-level comparison:** All metrics were aggregated by species. For wing regions, a Welch's t-test was performed to assess the statistical significance of mean intensity differences between species.
- **Visualization:** The distributions of wing and wingtip intensities, as well as the contrast metrics, were visualized using boxplots, violin plots, bar charts, line plots, and heatmaps to illustrate inter-species differences and the diagnostic value of each metric.

1.1.6 Implementation Details

All analyses were implemented in Python using the following libraries:

- **OpenCV:** Image loading, grayscale conversion, normalization, and region extraction
- **NumPy and Pandas:** Data manipulation and statistical calculations
- **SciPy:** Statistical testing (Welch's t-test)
- **Matplotlib and Seaborn:** Data visualization

1.1.7 Summary of Analytical Pipeline

1. Load each image and its segmentation mask.
2. Convert image to grayscale and normalize intensity.
3. Extract wing and wingtip pixels using color-coded masks.
4. Compute region-specific intensity statistics (mean, std, median, min, max, skewness, kurtosis).
5. Analyze wingtip pixel intensity distribution in 10-unit bins.
6. Quantify very dark pixels in the wingtip ($<30, 40, 50, 60$).
7. Calculate wing–wingtip contrast metrics at multiple thresholds.
8. Aggregate results by species and perform statistical tests.
9. Generate visualizations to support interpretability and species comparison.

2 Transfer Learning Approach

Transfer learning was employed in the implementation to leverage the robust feature extraction capabilities of pre-trained models on ImageNet. This approach aligns with best practices in fine-grained classification tasks, where lower-level features learned from diverse datasets can be effectively repurposed for specialized domains with limited data. The pre-training on ImageNet's 1.2 million images across 1,000 classes provides the model with a strong foundation for recognizing a wide range of visual patterns, which can then be fine-tuned for our specific classification task despite class imbalance challenges [Krizhevsky et al. \(2012\)](#).

Several pre-trained architectures were evaluated for this task, with VGG-16 [Simonyan and Zisserman \(2015\)](#) demonstrating superior performance in our specific classification context. The effectiveness of transfer learning was evident in the rapid convergence and high accuracy achieved even with our relatively limited dataset, demonstrating the potential of this approach for specialized classification tasks with significant class imbalance.

2.1 Common Implementation Strategy

All models except for the custom CNN utilized transfer learning to leverage knowledge from pre-trained networks. The transfer learning strategy included:

- Using models pre-trained on ImageNet as feature extractors
- Fine-tuning the entire network with a reduced learning rate (typically 0.0001 to 0.001)
- Replacing the final classification layer to output binary predictions (2 classes)
- Implementing dropout layers before final classification to prevent overfitting

This approach follows the established pattern that features learned in early layers of convolutional networks are more general and transferable, while later layers become more task-specific.

2.2 Data Preparation and Augmentation

Data augmentation was crucial to address the limited dataset size and class imbalance issues. Following best practices from [Cubuk et al.](#), multiple augmentation techniques were applied consistently across all models:

- **Spatial transformations:** Random horizontal flips, rotations (typically 15 degrees), and random/center crops were applied to increase geometric diversity.
- **Color space transformations:** Color jitter with brightness, contrast, and saturation adjustments of 0.2 magnitude was applied to make models robust to illumination variations.
- **Image enhancement:** In some implementations, sharpening filters were applied to improve feature clarity.
- **Normalization:** All images were normalized to match pre-trained model expectations [Shin et al.](#).

The augmentation strategy was deliberately more aggressive for the training set compared to validation and test sets, where only resizing, optional cropping, and normalization were applied to maintain evaluation consistency.

These techniques enhance model robustness to natural variations in image appearance, reducing overfitting and improving generalization capability [here](#).

2.3 Image Preprocessing

All images were preprocessed through a standardized pipeline:

Images were resized to match the architecture's expected input dimensions (224×224 pixels for most models, 299×299 pixels for Inception v3). Pixel values were normalized using ImageNet mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], ensuring input distributions aligned with those seen during pre-training [here](#).

2.4 Training Optimization Strategy

To optimize training with limited data, several techniques were employed consistently:

- **Optimizer:** AdamW optimizer with learning rates between 0.0001-0.001 and weight decay of 0.001-0.0005 was used across implementations to provide adaptive learning with regularization [here](#).
- **Learning rate scheduling:** Adaptive learning rate scheduling using either ReduceLROnPlateau or CosineAnnealingLR was implemented across models, reducing learning rates when validation metrics plateaued.
- **Early stopping:** Training was halted when validation accuracy stopped improving for a specified number of epochs (patience = 3-5) to prevent overfitting. [Early Stopping - But When?](#)

- **Gradient clipping:** Applied in some implementations to prevent gradient explosions and stabilize training. Due to the small and imbalanced dataset, gradient clipping was implemented to prevent limited images from causing large weight updates. [Why gradient clipping accelerates training: A theoretical justification for adaptivity. International Conference on Learning Representations \(ICLR\) here](#)
- **Loss function:** Cross-entropy loss was used consistently as the optimization objective for the binary classification task.
- **Mixed precision training:** For computationally intensive models like Inception V3, mixed precision training with torch.amp was used to improve computational efficiency.

The combination of these techniques enabled effective learning despite the challenges of limited data and class imbalance, with our best model achieving significantly better performance than traditional machine learning approaches on the same dataset.

2.5 Regularization Techniques

Multiple regularization strategies were employed to handle the limited data size and class imbalance:

- **Dropout:** Layers with rates between 0.3-0.4 were consistently added before final classification layers to reduce overfitting due to our small dataset size [Srivastava et al.](#).
- **Weight decay:** L2 regularization with weight decay values between 1e-4 and 1e-3 was applied across all models to prevent overfitting [Krogh & Hertz](#).
- **Batch normalization:** Used in custom CNN implementations to stabilize learning and improve convergence [Ioffe and Szegedy](#).
- **Data splitting:** Train/validation split of 80%/20% was consistently used to provide reliable validation metrics while maximizing training data.
- **Random seeds:** Fixed random seeds (42) were set for PyTorch, NumPy, and Python's random module to ensure reproducibility. Controlling randomness is essential for reliable hyperparameter tuning, performance assessment, and research reproducibility [here](#).

2.6 Addressing Class Imbalance

Our dataset exhibited significant class imbalance, which can degrade model performance by biasing predictions toward the majority class [here](#). To mitigate this challenge, multiple complementary strategies were implemented on the best performing models that included VGG16, and ViT:

- **Class-Weighted Loss Function**

- Implemented inverse frequency weighting (Cui et al., 2019) [\[link\]](#)

- Class weights calculation: $\text{class_weights}[i] = \frac{\text{total_samples}}{\text{num_classes} \times \text{label_counts}[i]}$

PyTorch implementation: `CrossEntropyLoss` with class weights tensor

- **Weighted Random Sampling**

- Balanced mini-batches using PyTorch's `WeightedRandomSampler`
- Sample weights: `samples_weights = class_weights[label]`
- Oversamples minority class and undersamples majority class [\[link\]](#)
- Uses replacement sampling for effective batch balancing

- **Class-Specific Data Augmentation**

- Aggressive minority class augmentation (Shorten & Khoshgoftaar, 2019) [\[link\]](#)
- Minority class transformations include:
 - * 30° random rotations
 - * Strong color jitter (brightness/contrast/saturation=0.3)
 - * Random resized crops (scale=0.7-1.0)
 - * Horizontal flips
 Standard augmentation for majority class (15° rotations, milder parameters)

2.7 Dataset Management

To address the challenges of limited data availability, an 80:20 train-validation split was implemented using random split stratification to maintain class distribution across partitions. This approach ensured that the validation set remained representative of the overall dataset while maximizing the samples available for training [Kohavi, 1995](#).

The batch size was set to 16, striking a balance between computational efficiency and optimization stability. Smaller batch sizes can increase gradient noise, which has been shown to act as an implicit regularizer that can improve generalization, particularly beneficial when working with limited training data [Keskar et al., 2016](#), [Masters & Luschi, 2018](#).

2.8 Evaluation Strategy

Model performance was systematically evaluated using:

- **Validation accuracy:** Used during training to select optimal model checkpoints and trigger early stopping or learning rate adjustments.
- **Test accuracy:** Final evaluation metric on the unseen test set to measure generalization performance.
- **Visualization:** Training loss and validation accuracy curves were plotted to analyze model convergence and potential overfitting.
- **Checkpointing:** Best-performing models based on validation accuracy were saved for later evaluation and deployment.

2.9 Model Checkpointing and Evaluation

Our implementation includes a robust evaluation framework with model checkpointing based on validation accuracy. This ensures that we preserve the best-performing model configuration throughout the training process. The model is trained for 20 epochs with early stopping implicitly implemented through best model saving. Performance is evaluated using accuracy on both validation and test sets, providing a comprehensive assessment of model generalization.

3 Model Architectures and Specific Implementations

3.1 VGG-16 Architecture

3.1.1 Theoretical Foundation

VGG-16 is a convolutional neural network architecture developed by Simonyan and Zisserman (2014) at the Visual Geometry Group (VGG) at Oxford, consisting of 16 weight layers including 13 convolutional layers followed by 3 fully connected layers. The architecture is characterized by its simplicity and depth, using small 3×3 convolutional filters stacked in increasing depth, followed by max pooling layers. With approximately 138 million parameters, VGG-16 provides a strong foundation for feature extraction in computer vision tasks.

The primary advantage of employing VGG-16 for transfer learning in fine-grained classification tasks is its hierarchical feature representation capability, which enables the capture of both low-level features (edges, textures) and high-level semantic features. Pre-trained on the ImageNet dataset containing over 1.2 million images across 1,000 classes, VGG-16 offers robust initialization weights that facilitate effective knowledge transfer to domain-specific tasks with limited training data.

VGG-16 has demonstrated superior performance in fine-grained classification tasks compared to conventional techniques. Recent studies show that VGG-16 with logistic regression achieved 97.14% accuracy on specialized datasets like Leaf12, significantly outperforming traditional approaches that combined color channel statistics, texture features, and classic classifiers which only reached 82.38% accuracy [here](#). For our specific task of gull species classification, the hierarchical feature representation capabilities of VGG-16 proved particularly effective at capturing the subtle differences in wing patterns and morphological features that distinguish between the target species.

3.1.2 Model Adaptation for Fine-Grained Classification

For our specific fine-grained binary classification task with limited data and class imbalance, the VGG-16 architecture was adapted through a targeted modification strategy:

- The pre-trained VGG-16 model was loaded with ImageNet weights.
- The feature extraction layers (convolutional base) were preserved to maintain the rich hierarchical representations learned from ImageNet.
- The original 1000-class classifier was replaced with a custom binary classification head consisting of:

- A dropout layer with a rate of 0.4 to reduce overfitting.
- A fully-connected layer mapping from the original 4096 features to 2 output classes.

(?) demonstrated that VGG-16 achieves 94.3% accuracy on CUB-200-2011 by fine-tuning only the final three layers, a strategy mirrored in my VGG implementation where the classifier head was replaced while preserving ImageNet-initialized convolutional weights. This approach aligns with successful methodologies in avian species classification using VGG-16 as demonstrated by Brown et al. (2018), where fine-tuning the architecture by modifying the final classification layer enabled the model to retain general feature recognition capabilities while adapting to species-specific visual characteristics [here](#).

3.2 Vision Transformer (ViT) Architecture

3.2.1 ViT for Fine-Grained Classification

Vision Transformers (ViT) have emerged as powerful alternatives to convolutional neural networks for visual recognition tasks. First introduced by Dosovitskiy et al. ([Dosovitskiy et al., 2021](#)), ViTs process images as sequences of fixed-size patches, applying transformer-based self-attention mechanisms to model global relationships between image regions. This architecture enables the capture of long-range dependencies within images, making it particularly suitable for fine-grained classification tasks where subtle distinctions between similar classes may depend on relationships between distant image features.

3.2.2 Vision Transformer Implementation

For our primary approach, a Vision Transformer using transfer learning from a pre-trained model was implemented:

- Base architecture: 'vit_base_patch16_224' pre-trained on ImageNet from the TIMM library ([Wightman, 2021](#))
- Input resolution: 224×224 pixels with 16×16 pixel patches
- Feature dimension: 768-dimensional embeddings
- Adaptation strategy: Replacement of the classification head with a binary classifier while preserving the pre-trained transformer blocks

The model architecture preserves the core self-attention mechanism of ViT while adapting the final classification layer for our specific binary classification task. This approach follows established transfer learning principles for vision transformers ([Touvron et al., 2021](#)), leveraging representations learned from large-scale datasets to overcome our limited training data constraints.

3.2.3 Alternative ViT Implementations

In addition to our primary implementation, we explored two attention-enhanced architectures:

InterpretableViT We developed an InterpretableViT model that incorporates explicit attention mechanisms for improved focus on discriminative features:

- Separates the class token from patch tokens
- Applies a learned attention layer to generate importance weights for each patch
- Combines the class token with attention-weighted patch representations
- Employs a multi-layer classifier with dropout regularization

A key advantage of this architecture is its compatibility with gradient-based visualization techniques. By separating the class token from patch tokens and implementing an explicit attention mechanism, the model facilitates more effective application of Grad-CAM ([Selvaraju et al., 2017](#)), allowing for visualization of discriminative image regions contributing to classification decisions.

EnhancedViT We also implemented an EnhancedViT that applies attention-based weighting across all tokens:

- Processes all tokens (including class token) through an attention mechanism
- Generates a single attention-weighted feature representation
- Utilizes a specialized classification head with dropout for regularization

This implementation draws from research on token aggregation strategies in vision transformers ([Wang et al., 2021](#)), which shows that attention-weighted token aggregation can improve performance in data-limited regimes.

3.3 Inception v3 Architecture

3.3.1 Theoretical Background

Inception v3, developed by Szegedy et al. (2016), represents a sophisticated CNN architecture designed to efficiently capture multi-scale features through parallel convolution pathways with varied kernel sizes. The key innovation in Inception architectures is the utilization of *Inception modules* that process the same input tensor through multiple convolutional paths with different receptive fields, and then concatenate the results. This enables the network to capture both fine-grained local patterns and broader contextual information simultaneously ([Szegedy et al., 2016](#)).

Inception v3 builds upon earlier versions with several important architectural improvements:

- Factorized convolutions to reduce computational complexity
- Spatial factorization into asymmetric convolutions (e.g., $1 \times n$ followed by $n \times 1$)
- Auxiliary classifiers that inject additional gradient signals during training
- Batch normalization for improved training stability and faster convergence

- Label smoothing regularization to prevent overconfidence

These design elements collectively enable Inception v3 to achieve high accuracy while maintaining computational efficiency. As demonstrated by Huang et al. (2019), Inception architectures are particularly effective for tasks requiring multi-scale feature extraction, such as discriminating between visually similar biological specimens (Huang et al., 2019).

3.3.2 Model-Specific Implementation Details

Our implementation adapted the pre-trained Inception v3 model for fine-grained gull species classification with the following specific elements:

- **Input Resolution:** Resize operations were performed to 299×299 pixels (the standard input size for Inception v3). The larger input resolution (299×299 vs 224×224 used by VGG16) provides the Inception architecture with more detailed information, potentially beneficial for capturing the subtle wing pattern differences between gull species (Xie et al., 2020).
- **Auxiliary Outputs:** A distinctive aspect of our Inception v3 implementation was the utilization of auxiliary outputs during training. Inception v3's auxiliary classifier, which branches off from an intermediate layer, provides an additional gradient path during backpropagation. This helps combat the vanishing gradient problem and provides regularization with auxiliary loss weight of 0.3 (He et al., 2019).
- **Mixed-Precision Training:** We employed PyTorch's Automatic Mixed Precision (AMP) to accelerate computation while maintaining numerical stability (Micikevicius et al., 2018). This technique allows the use of float16 precision where appropriate, which reduces memory usage and increases computational speed, especially beneficial when training on GPU-constrained environments like Google Colab.

3.4 Residual Network (ResNet-50) Implementation

Residual Networks (ResNet) have revolutionized deep learning architectures by introducing identity shortcut connections that bypass one or more layers, enabling the training of substantially deeper networks He et al., 2016. These skip connections address the degradation problem by allowing gradients to flow more effectively during backpropagation, mitigating the vanishing gradient issue prevalent in very deep neural networks.

For our fine-grained gull species classification task, a transfer learning approach based on the ResNet-50 architecture was implemented. This implementation was motivated by ResNet's demonstrated success in capturing hierarchical features at multiple levels of abstraction, which is particularly valuable for distinguishing the subtle morphological differences between visually similar gull species He et al., 2016b, Zhao et al., 2019.

3.4.1 Architecture-Specific Enhancements

A distinctive aspect of our ResNet-50 implementation was the incorporation of image sharpening as a preprocessing technique. We applied a 3×3 Laplacian sharpening kernel to enhance edge definition and accentuate the subtle diagnostic features crucial for distinguishing between gull species

[Gonzalez & Woods, 2018](#). This approach was inspired by research showing that edge enhancement can improve the detection of fine-grained morphological features in avian classification tasks [Berg et al., 2014](#).

The sharpening kernel was systematically applied to both training and evaluation pipelines using OpenCV's `filter2D` function, ensuring consistent feature enhancement across all dataset partitions. This preprocessing step proved particularly valuable for highlighting distinctive wingtip patterns and subtle plumage characteristics that serve as key discriminative features between the target species [Dutta & Zisserman, 2019](#).

3.5 Custom CNN with Squeeze-and-Excitation Blocks

3.5.1 Architectural Design

To address the challenges of limited data and class imbalance in fine-grained classification, we developed a lightweight custom CNN architecture incorporating attention mechanisms. Our approach employs Squeeze-and-Excitation (SE) blocks, which enhance feature representation by modeling channel interdependencies through an attention mechanism. The SE block, as introduced by Hu et al. ([2018](#)), adaptively recalibrates channel-wise feature responses to emphasize informative features while suppressing less useful ones.

The architecture consists of three convolutional blocks, each followed by batch normalization, ReLU activation, and an SE block. The SE block performs two key operations:

- **Squeeze:** Global average pooling across spatial dimensions to generate channel-wise statistics
- **Excitation:** A fully connected layer that produces modulation weights for each channel

This channel-wise attention mechanism has been shown to improve model performance with minimal computational overhead ([Hu et al., 2018](#)). The SE blocks in our implementation use a reduction ratio of 16, balancing parameter efficiency and representational power.

3.5.2 Custom CNN-Specific Training Approach

Unlike the transfer learning approaches used with pre-trained models, our custom CNN was trained from scratch with some specific optimization strategies:

- **Cosine Annealing scheduler:** Our learning rate schedule follows a cosine annealing pattern with a period of 10 epochs, allowing the learning rate to oscillate and potentially escape local minima ([Loshchilov and Hutter, 2017](#)).
- **Specialized augmentation:** The custom CNN particularly benefited from more aggressive data augmentation strategies to compensate for the lack of pre-trained weights, including stronger rotations and more extensive color jittering than used with the transfer learning models.

Algorithm 1 VGG16Modified Architecture

```
1: function VGG16MODIFIED
2:   Load pre-trained VGG-16 with ImageNet weights
3:   Extract number of features from final layer:  $num\_ftrs \leftarrow \text{VGG.classifier}[6].in\_features$ 
4:   Replace final classifier layer with:
5:   Dropout( $p=0.4$ )
6:   Linear( $num\_ftrs \rightarrow 2$ ) ▷ Binary classification
7: end function
8: function FORWARD( $x$ )
9:   return VGG( $x$ )
10: end function
```

Algorithm 2 ViTModified Architecture

```
1: function ViTMODIFIED
2:   Load pre-trained ViT (vit_base_patch16_224) with ImageNet weights
3:   Extract number of features:  $num\_ftrs \leftarrow \text{ViT.head.in\_features}$ 
4:   Replace classification head:  $\text{ViT.head} \leftarrow \text{Linear}(num\_ftrs \rightarrow 2)$ 
5: end function
6: function FORWARD( $x$ )
7:   return ViT( $x$ )
8: end function
```

Algorithm 3 EnhancedViT Architecture

```
1: function ENHANCEDViT(dropout_rate=0.3, hidden_dim=512)
2:   Load pre-trained ViT (vit_base_patch16_224) with ImageNet weights
3:   Remove classification head:  $\text{ViT.head} \leftarrow \text{Identity}()$ 
4:   Extract embedding dimension:  $embed\_dim \leftarrow \text{ViT.embed\_dim}$ 
5:   Create attention layer:
6:   Linear( $embed\_dim \rightarrow 1$ )
7:   Create classifier:
8:   LayerNorm( $embed\_dim$ )
9:   Dropout( $dropout\_rate$ )
10:  Linear( $embed\_dim \rightarrow hidden\_dim$ )
11:  ReLU()
12:  Dropout( $dropout\_rate$ )
13:  Linear( $hidden\_dim \rightarrow 2$ )
14: end function
15: function FORWARD( $x$ )
16:    $tokens \leftarrow \text{ViT.forward\_features}(x)$  ▷ Shape: [batch, num_tokens, embed_dim]
17:    $attn\_scores \leftarrow \text{attention\_layer}(tokens)$  ▷ Shape: [batch, num_tokens, 1]
18:    $attn\_weights \leftarrow \text{Softmax}(attn\_scores, \text{dim}=1)$ 
19:    $weighted\_feature \leftarrow \text{Sum}(attn\_weights \cdot tokens, \text{dim}=1)$ 
20:    $out \leftarrow \text{classifier}(weighted\_feature)$ 
21:   return  $out$ 
22: end function
```

Algorithm 4 InterpretableViT Architecture

```
1: function INTERPRETABLEViT(dropout_rate=0.3, hidden_dim=512)
2:   Load pre-trained ViT (vit_base_patch16_224) with ImageNet weights
3:   Remove classification head: ViT.head  $\leftarrow$  Identity()
4:   Extract embedding dimension:  $embed\_dim \leftarrow$  ViT.embed_dim
5:   Create attention layer for patch tokens:
6:   Linear( $embed\_dim \rightarrow 1$ )
7:   Create classifier for combined representation:
8:   LayerNorm( $embed\_dim \cdot 2$ )
9:   Dropout(dropout_rate)
10:  Linear( $embed\_dim \cdot 2 \rightarrow hidden\_dim$ )
11:  ReLU()
12:  Dropout(dropout_rate)
13:  Linear( $hidden\_dim \rightarrow 2$ )
14: end function
15: function FORWARD( $x$ )
16:   $tokens \leftarrow$  ViT.forward_features( $x$ ) ▷ Shape: [batch, N+1, embed_dim]
17:   $cls\_token \leftarrow tokens[:, 0, :]$  ▷ Extract CLS token
18:   $patch\_tokens \leftarrow tokens[:, 1 :, :]$  ▷ Extract patch tokens
19:   $attn\_scores \leftarrow$  attention_layer(patch_tokens)
20:   $attn\_weights \leftarrow$  Softmax(attn_scores, dim=1)
21:   $weighted\_patch \leftarrow$  Sum(attn_weights  $\cdot$  patch_tokens, dim=1)
22:   $combined \leftarrow$  Concatenate(cls_token, weighted_patch, dim=1)
23:   $logits \leftarrow$  classifier(combined)
24:  return logits, attn_weights ▷ Return both for interpretability
25: end function
```

Algorithm 5 InceptionV3Modified Architecture

```
1: function INCEPTIONV3MODIFIED
2:   Load pre-trained Inception v3 with ImageNet weights
3:   Extract number of features:  $num\_ftrs \leftarrow$  Inception.fc.in_features
4:   Replace classification layer:
5:   Dropout(0.5)
6:   Linear( $num\_ftrs \rightarrow 2$ ) ▷ Binary classification
7: end function
8: function FORWARD( $x$ )
9:   if self.training then
10:     $aux\_out, out \leftarrow$  Inception( $x$ ) ▷ Return auxiliary output during training
11:    return aux_out, out
12:   else
13:    return Inception( $x$ )
14:   end if
15: end function
```

Algorithm 6 ResNet50Modified Architecture

```
1: function RESNET50MODIFIED
2:   Load pre-trained ResNet-50 with ImageNet weights
3:   Extract number of features:  $num\_ftrs \leftarrow ResNet.fc.in\_features$ 
4:   Replace final fully connected layer:
5:   Dropout(0.5)
6:   Linear( $num\_ftrs \rightarrow 2$ ) ▷ Binary classification
7: end function
8: function FORWARD( $x$ )
9:   return ResNet( $x$ )
10: end function
```

Algorithm 7 Custom CNN with Squeeze-and-Excitation Blocks

```
1: function SEBLOCK(channels, reduction=16)
2:   Create fully connected layers:
3:    $fc1 \leftarrow \text{Linear}(\text{channels} \rightarrow \text{channels}/\text{reduction})$ 
4:    $fc2 \leftarrow \text{Linear}(\text{channels}/\text{reduction} \rightarrow \text{channels})$ 
5: end function
6: function SEBLOCK.FORWARD( $x$ )
7:    $\text{batch}, \text{channels}, \leftarrow x.\text{size}()$ 
8:    $se \leftarrow \text{Mean}(x, \text{dims}=(2, 3))$  ▷ Global Average Pooling
9:    $se \leftarrow \text{ReLU}(fc1(se))$ 
10:   $se \leftarrow \text{Sigmoid}(fc2(se))$ 
11:   $se \leftarrow se.\text{view}(\text{batch}, \text{channels}, 1, 1)$ 
12:  return  $x \cdot se$  ▷ Channel-wise multiplication
13: end function
14: function IMPROVEDCNN
15:   Create convolutional layers:
16:   Conv2d( $3 \rightarrow 32$ , kernel=3, stride=1, padding=1)
17:   BatchNorm2d(32)
18:   ReLU()
19:   SEBlock(32)
20:   MaxPool2d(kernel=2, stride=2)
21:   Conv2d( $32 \rightarrow 64$ , kernel=3, stride=1, padding=1)
22:   BatchNorm2d(64)
23:   ReLU()
24:   SEBlock(64)
25:   MaxPool2d(kernel=2, stride=2)
26:   Conv2d( $64 \rightarrow 128$ , kernel=3, stride=1, padding=1)
27:   BatchNorm2d(128)
28:   ReLU()
29:   SEBlock(128)
30:   MaxPool2d(kernel=2, stride=2)
31:   Create fully connected layers:
32:   Flatten()
33:   Linear( $128 \cdot 16 \cdot 16 \rightarrow 512$ )
34:   ReLU()
35:   Dropout(0.5)
36:   Linear( $512 \rightarrow 2$ ) ▷ Binary classification
37: end function
38: function IMPROVEDCNN.FORWARD( $x$ )
39:    $x \leftarrow \text{conv\_layers}(x)$ 
40:    $x \leftarrow \text{fc\_layers}(x)$ 
41:   return  $x$ 
42: end function
```

Algorithm 8 Transfer Learning Framework for Fine-Grained Classification

```
1: function TRANSFERLEARNINGMODEL(model_name, num_classes=2)
2:   Load pre-trained model with ImageNet weights
3:   Extract model-specific features:
4:   Identify final layer dimensions
5:   Preserve feature extraction layers/blocks
6:   Replace classification head:
7:   Apply regularization (e.g., Dropout with appropriate rate)
8:   Create new classifier mapping to target classes
9:   Configure optimization strategy:
10:  Select appropriate optimizer (Adam/AdamW)
11:  Set learning rate and weight decay parameters
12:  Implement learning rate scheduler
13:  Define data augmentation pipeline:
14:  Apply standard transformations (resize, normalize)
15:  Implement model-specific augmentations
16:  Return modified model
17: end function
18: function MODELFORWARD(input)
19:   Extract features from modified backbone
20:   Apply classification head
21:   Return predictions
22: end function
23: function TRAININGLOOP(model, train_loader, val_loader, epochs)
24:   for epoch in 1 to epochs do
25:     Train model on training data
26:     Evaluate on validation data
27:     Update learning rate based on scheduler
28:     Save best model based on validation metrics
29:   end for
30: end function
```
