

Article

Fine-Grained Soybean Variety Recognition Using Swin Transformer with Differential Attention and Dynamic Channel Aggregation

Puyuan Shi ¹, Qilin Yang ¹, Liuchao Zhu ¹, Zixin Chen ¹, Huanliang Xu ¹, Ji Huang ^{2,*} and Junxian Huang ^{1,*} 

¹ College of Artificial Intelligence, Nanjing Agricultural University, No. 1 Weigang, Nanjing 210095, China; 2024819034@stu.njau.edu.cn (P.S.); 2024819036@stu.njau.edu.cn (Q.Y.); 2023119003@stu.njau.edu.cn (L.Z.); czx@stu.njau.edu.cn (Z.C.); huanliangxu@njau.edu.cn (H.X.)

² College of Agriculture, Nanjing Agricultural University, No. 1 Weigang, Nanjing 210095, China

* Correspondence: huangji@njau.edu.cn (J.H.); jim@njau.edu.cn (J.H.)

Abstract

Soybean variety recognition supports germplasm management and intelligent agricultural inspection, but remains challenging because varieties often show subtle inter-class differences, large intra-class imaging variation, and imbalanced samples. This study adopts a cascaded pipeline that combines YOLO-based object localization, manual verification and cropping, and fine-grained classification, allowing the classifier to learn from standardized cropped soybean seed images rather than original whole images. To improve Swin Transformer for this task, we propose Swin-Diff-DCA, which introduces Dynamic Channel Aggregation (DCA) in Stage 3 for middle-level local feature reuse and a differential attention branch in Stage-4 window attention for deep discriminative enhancement. On a dataset containing 25 soybean varieties and 1511 test images, Swin-Diff-DCA achieves average results of 87.18% Accuracy, 82.33% Macro-F1, and 87.44% Weighted-F1 across three random seeds under the current split and evaluation protocol. It outperforms RegNet, ResNet-50, ViT, the Swin Transformer baseline, and single-module variants. The results show that combining middle-level feature reuse with deep differential enhancement improves cropped soybean variety classification, while low-sample and visually similar classes remain the main sources of misclassification.

Keywords: soybean variety recognition; fine-grained image classification; Swin Transformer; differential attention; Dynamic Channel Aggregation; agricultural image analysis



Academic Editor: Juan Ignacio Arribas

Received: 11 June 2026

Revised: 10 July 2026

Accepted: 10 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Soybean variety recognition is an important task in germplasm resource management, breeding material screening, and seed quality inspection. Accurate identification of different soybean varieties can help improve the efficiency of sample archiving, variety verification, and agricultural production supervision. Traditional recognition methods usually rely on manual experience or physicochemical testing. Although these methods have certain reliability, they are easily limited by detection efficiency, labor cost, and subjective judgment when the sample size is large, the number of varieties is high, or rapid detection is required. Therefore, constructing an image-based automatic soybean variety recognition method has clear research significance and application value, consistent with recent work on high-throughput seed phenotyping using deep learning [1]. Machine-learning-based soybean seed and seedling quality classification has also shown the potential of

replacing labor-intensive manual inspection with data-driven visual assessment. Existing soybean seed recognition studies have explored several visual and spectral analysis strategies for variety identification and seed quality assessment. Hyperspectral imaging and spectroscopy-based methods provide rich spectral information and have shown effectiveness in distinguishing seed varieties or defects, but they usually depend on specialized acquisition equipment and may be less convenient for low-cost image-based recognition scenarios [2–5]. Transfer-learning, ensemble-learning, attention-based, and deep neural network methods further demonstrate that data-driven visual features can improve soybean seed classification performance compared with purely manual inspection [2–5]. These studies provide an important basis for automatic soybean seed recognition, while the present study focuses on fine-grained classification using ordinary RGB images after target localization and manual cleaning, where the key challenge is to distinguish visually similar soybean varieties from subtle color, texture, contour, and local morphological differences.

This paper adopts a two-stage process of object detection and image classification to complete soybean variety recognition. First, a YOLO object detection model is used to locate soybean seed regions in the original images [6], and single-seed or target soybean seed images are cropped according to the detection boxes. Then, the cropped seed images are fed into a deep learning classification model to identify the specific variety. Compared with directly classifying the whole original image, this process can reduce the interference caused by background, irrelevant regions, and target scale variation, allowing the subsequent classification stage to focus more on discriminative features of soybean seeds, such as color, shape, contour, texture, and local color differences.

Even after YOLOv11 detection [7] and target cropping, soybean variety classification still has obvious fine-grained recognition difficulty. Different soybean varieties are highly similar in overall appearance, and inter-class differences are often concentrated in local texture, seed coat color, contour details, and subtle morphological changes. At the same time, images within the same variety may also be affected by shooting angle, lighting conditions, seed pose, and individual differences, resulting in large intra-class variation in the cropped seed images. In addition, the number of samples among different varieties is imbalanced in the actual dataset, and the model may tend to favor categories with more samples, thereby affecting the recognition performance of minority varieties. These factors jointly make soybean seed variety recognition a challenging long-tailed fine-grained classification task.

Existing deep learning classification models provide important architectural foundations for image-based soybean seed recognition. CNN-based models such as ResNet and RegNet are effective in extracting hierarchical local visual patterns, and their residual or network-design strategies have been widely used to improve feature representation in image classification tasks [8,9]. Transformer-based models were first established through attention-based sequence modeling [10] and were later extended to image recognition through Vision Transformer, which demonstrates the potential of self-attention for modeling global relationships among image patches [11]. Swin Transformer further introduces hierarchical representation and shifted-window attention, making it suitable for image recognition tasks that require both local modeling and multi-scale feature representation [12]. These models provide useful baselines for soybean seed classification because they can represent local texture, shape, and broader spatial relationships in cropped seed images. However, directly using the Swin Transformer backbone may still create two problems in this task: first, middle-level local structural information may be weakened during layer-by-layer propagation; and second, deep window attention may be insufficient to highlight fine-grained discriminative differences during semantic aggregation. Therefore, the structural improvements in this study mainly focus on two levels: enhancing the trans-

mission of local texture, contour, and color-boundary information in the middle stage, and strengthening the response to weak difference regions between similar varieties in the deep semantic stage.

This paper proposes a Swin-Diff-DCA classification model for soybean seed variety recognition, which is used in the variety classification stage of the two-stage process. The model uses Swin Transformer as the backbone. A differential attention branch, inspired by differential attention modeling, is introduced into Stage-4 window attention to enhance deep discriminative feature representation through the differential response between auxiliary attention and original attention. Meanwhile, Dynamic Channel Aggregation (DCA) is introduced in Stage 3 and implemented as a dynamic residual channel routing mechanism. It uses intermediate features of anchor blocks to provide channel-level residual supplementation for target blocks, thereby strengthening the transmission of middle-level local structural information to deep semantic features. On the basis of preserving the representation ability of the pretrained backbone, this design improves the model's sensitivity to local discriminative differences required for fine-grained soybean seed classification.

To improve training stability, the training process combines ImageNet pretrained weights, label smoothing, grouped learning rates, freeze warm-up, cosine learning-rate scheduling, mixed-precision training, AdamW optimization, and gradient clipping. In the evaluation stage, Accuracy, Weighted-F1, and Macro-F1 are used comprehensively. Weighted-F1 reflects the overall recognition performance under the sample distribution of the test set, while Macro-F1 is used to assist in observing the balanced classification performance among different varieties.

The main scientific contributions of this study are threefold. First, this study constructs and organizes a cropped-and-cleaned RGB soybean seed image dataset for fine-grained variety classification. The dataset construction process explicitly separates YOLO-based object localization, manual sample verification, and classification-stage evaluation, which helps reduce background interference and clarifies that the reported performance mainly reflects cropped-image classification rather than complete end-to-end deployment performance. Second, this study proposes Swin-Diff-DCA, a Swin Transformer-based classification model designed for visually similar soybean varieties. Different from directly using a general Swin Transformer backbone, the proposed model introduces Dynamic Channel Aggregation (DCA) in Stage 3 to strengthen the reuse of middle-level local texture, contour, and color-boundary information and introduces a differential attention branch in Stage 4 to enhance deep discriminative responses to weak inter-class differences. Third, under a unified data split and evaluation protocol, this study evaluates the proposed model through baseline comparisons, ablation experiments, ViT-based supplementary experiments, confusion matrix analysis, and Grad-CAM visualization. These experiments provide evidence for the effectiveness and limitations of the proposed modules under the current dataset setting.

The rest of this paper is organized as follows. Section 2 introduces dataset construction, image preprocessing, model structure, evaluation metrics, and implementation details. Section 3 presents experimental results, including model comparison, ablation experiments, ViT-based supplementary experiments, and confusion matrix analysis. Section 4 discusses metric interpretation, model effectiveness, interpretability, and limitations. Section 5 summarizes the study and outlines future work.

2. Materials and Methods

2.1. Materials

2.1.1. Data Collection

This study focuses on fine-grained visual classification of soybean varieties and constructs an image dataset for variety recognition. The original images were collected from

June to November 2025 from soybean seeds of different varieties, and the category labels were organized according to the actual variety names. Image acquisition was conducted under relatively stable indoor conditions using a Sony ZV-E10 digital camera and an iPhone 15 Pro smartphone. The original images captured by the digital camera had a resolution of 6000×4000 pixels, while the images captured by the smartphone had a resolution of 3024×4032 pixels. During image acquisition, soybean seeds were placed on both white and black backgrounds, and relatively stable indoor illumination was used to maintain consistent lighting conditions. Since different soybean varieties often differ only in subtle aspects such as color, texture, markings, contour, and local morphology, while images within the same variety may also be affected by illumination, shooting angle, seed pose, and individual differences, this task can be regarded as a typical fine-grained visual classification problem [13,14].

In this study, classification was not performed directly on the entire original images. Instead, this study adopts a cascaded data construction process of YOLO-based object localization, manual verification and cropping, and fine-grained variety classification. First, a YOLO object detection model is used to locate soybean seeds or valid target regions in the original images. Then, candidate regions are cropped according to the detection boxes, and samples with false detections, severe occlusion, blur, incomplete targets, or uncertain labels are removed through manual verification. Finally, the cleaned cropped images are organized by variety category to form a standard image classification dataset. This process can reduce the interference of complex backgrounds and non-target regions, allowing the subsequent model to focus more on the discriminative features of soybean seeds themselves. The data volume of each soybean variety is shown in Table 1. It should be noted that augmented images are only used for training-stage data expansion, and the validation and test sets do not contain randomly augmented images.

Table 1. Category data volume.

Variety	Original Images	Manually Cropped Images	Training Augmented Images	Total Images
nn49	28	967	1456	2423
nn60	28	366	590	956
zld105	36	357	628	985
sn29	52	578	844	1422
lk314	38	742	1100	1842
nn47	31	450	616	1066
jd17	28	250	524	774
sd30	124	929	1444	2373
hd16	62	479	736	1215
jng20839	62	468	800	1268
nn43	78	596	888	1484
nn42	60	496	710	1206
sn23	66	503	826	1329
b73	64	610	1038	1648
sz2	59	485	850	1335
zd57	59	342	524	866
xd18	63	544	954	1498
zd53	59	450	750	1200
zd61	55	484	746	1230
zd59	55	317	528	845
zd51	47	274	492	766
sd29	45	227	384	611
xzd1	169	1469	2308	3777
zh301	186	1010	1552	2562
nn55	181	1240	1896	3136

Table 1 reports the candidate classification images generated during manual cropping and training-set augmentation, and it is not equivalent to the final number of images used for training, validation, and testing. After quality review, sample cleaning, and data splitting, the final dataset used for classification model training and evaluation contains 37,317 images, including 34,251 training images, 1555 validation images, and 1511 test images. The augmented images are used only for model training, while the validation and test sets do not contain augmented copies.

2.1.2. Dataset Splitting

The dataset is divided into training, validation, and test sets. The training set is used for model parameter learning, the validation set is used for model selection and hyperparameter adjustment, and the test set is used only for final performance evaluation. The split is performed within each category, and a fixed random seed is used to ensure that different models are trained and compared under exactly the same data split. To avoid data leakage, samples from the same original image or highly similar cropped regions are not assigned to both the training and test sets. The resulting split is summarized in Table 2.

Table 2. Training/validation/test split.

Dataset Split	Number of Images	Augmentation Used	Purpose
Training Set	34,251	Yes	Parameter Learning
Validation Set	1555	No	Model Selection
Test Set	1511	No	Final Evaluation

Since the soybean variety dataset has class imbalance, using Accuracy alone as the model selection metric may bias the model toward categories with more samples [15]. Therefore, this paper uses Macro-F1 as the core metric for model selection and experimental analysis, while also reporting Accuracy, Precision, Recall, Weighted-F1, and average loss to evaluate the model more comprehensively on long-tailed and difficult categories.

To reduce the risk of sample leakage, dataset splitting was performed at the original-image level for source-related cropped samples. Specifically, all cropped samples generated from the same original image were assigned to the same subset, namely the training, validation, or test set, so that source-related crops did not appear simultaneously across different subsets. During sample cleaning, near-duplicate or highly similar crops were also checked to reduce overlap among the training, validation, and test sets. This strategy was used to reduce the possibility of inflated test performance caused by similar samples appearing in different subsets. The overall cascaded process is illustrated in Figure 1.

2.2. Data Preprocessing

This section introduces the preprocessing pipeline from original images to classification model inputs, including YOLO object localization, manual verification and cropping, image standardization, and data augmentation. It should be noted that the detection stage is used to obtain classification input regions, while the main evaluation target of this study is the classification model on cropped and cleaned images.

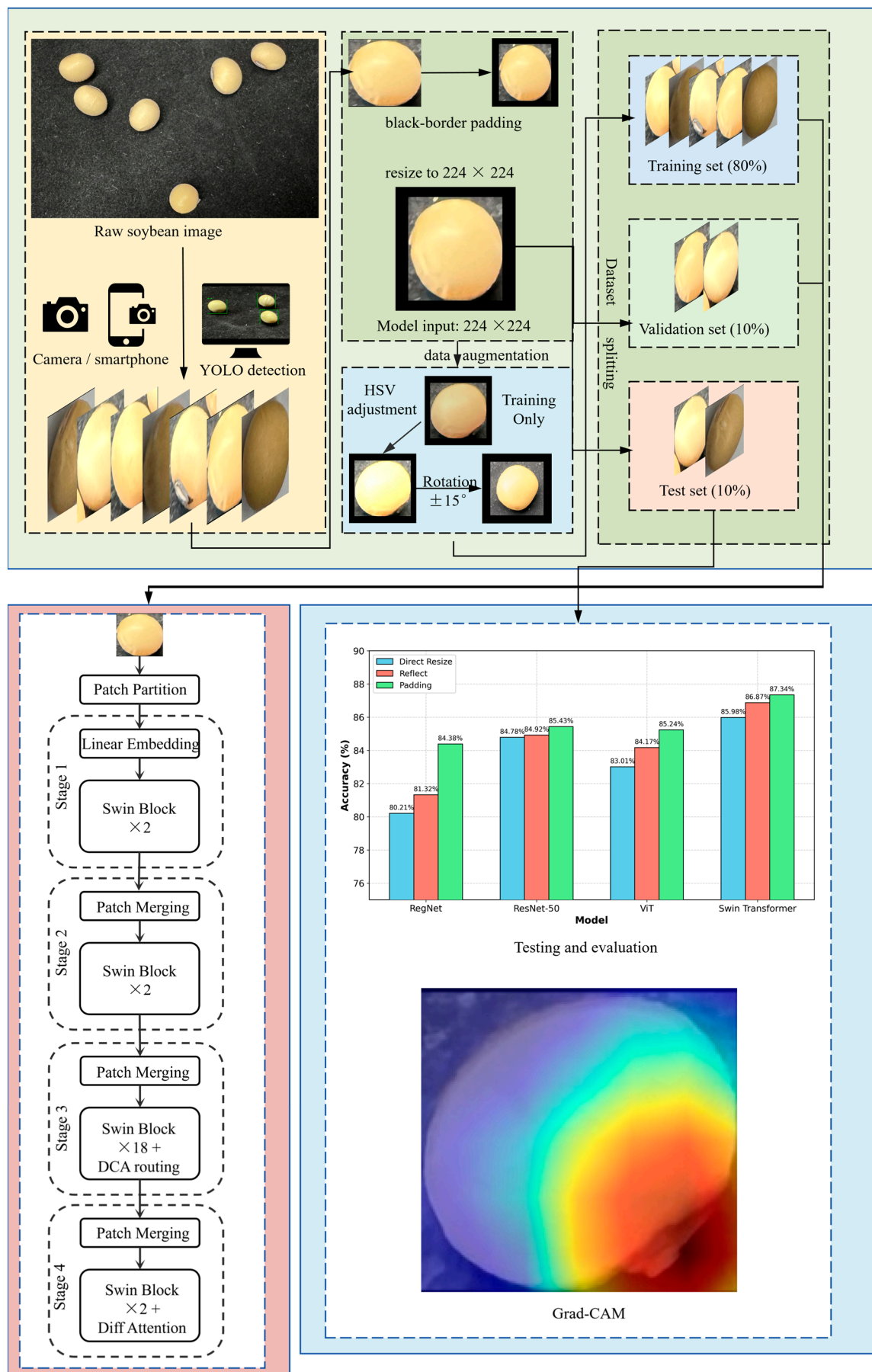


Figure 1. Overall pipeline of soybean variety recognition.

2.2.1. Object Detection

In the preprocessing stage, a YOLO object detection model is first used to locate targets in the original soybean images. Given an original image, the detection model outputs candidate bounding boxes and their confidence scores, which are used to determine the locations of soybean seeds or valid local regions in the image. Candidate bounding boxes with confidence scores lower than 0.25 were discarded, and non-maximum suppression with an IoU threshold of 0.60 was applied to remove duplicated detections. The remaining bounding boxes were used as the initial cropping regions for classification-stage image construction. The purpose of this step is not to directly complete variety classification, but to obtain more stable input regions for the classification stage from complex original images.

After automatic detection, the candidate cropped images are further manually verified. Samples with shifted detection boxes, incomplete targets, severe blur, obvious occlusion, or uncertain category labels are removed or corrected. After this step, the classification stage uses soybean images that have been cropped, cleaned, and organized by category, rather than the original whole images. This processing strategy decouples the detection task from the classification task, so that the classification model performance mainly reflects fine-grained variety recognition ability rather than object localization quality. The YOLO version used in this paper is YOLOv11. The detection targets are soybean seeds or valid local regions. The cropping regions are determined by detection boxes. Manual verification mainly removes false detections, severe occlusion, blurred images, incomplete targets, and samples with uncertain labels. The subsequent experiments mainly evaluate classification model performance on cropped and cleaned images, and YOLO detection accuracy is not used as the core evaluation object. The YOLO data construction settings are summarized in Table 3.

Table 3. YOLO data construction settings.

Item	Setting
Detection model	YOLOv11
Detection target	Soybean seeds or valid target regions
Input images	Raw collected images
Cropping basis	YOLO detection boxes
Cropping strategy	Crop according to detection boxes, apply boundary correction when necessary, then resize proportionally and pad with black borders
Manually removed samples	False detections, severe occlusion, blurred images, incomplete targets, and samples with uncertain labels
Classification-stage input	Images manually verified, cropped, cleaned, and organized by category
Confidence Threshold	0.25
IoU threshold for NMS	0.6
Manual verification criteria	False detections, shifted boxes, incomplete seeds, severe occlusion, blur, strong shadow interference, and uncertain labels

2.2.2. Image Standardization and Data Augmentation

The target regions cropped by YOLO vary in aspect ratio and scale, so they need to be standardized to a fixed input size. Directly stretching the cropped images to a fixed size is simple, but it changes the real geometric shape of soybean seeds. For soybean variety recognition, seed contour, length-width ratio, edge morphology, and local texture distribution may all contain important discriminative information. If non-proportional stretching is used, these fine-grained structural cues may be artificially distorted.

To preserve the original aspect ratio of the target region, this paper adopts image standardization based on proportional resizing and black-border padding. Specifically, the original aspect ratio of the cropped image is first maintained, and the longer side is resized to the target scale. Then, black pixels are padded along the shorter side so that the image becomes a fixed-size square input. This method satisfies the fixed input size requirement of Swin Transformer while avoiding geometric deformation caused by direct stretching.

Aspect-ratio-preserving padding was used to avoid geometric distortion of seed contours and local morphological cues. Compared with direct stretching, proportional resizing with black padding preserves the original seed shape while producing fixed-size square inputs required by the classification backbone. The comparison between direct stretching and aspect-ratio-preserving padding is shown in Figure 2.

Effect of Padding on Accuracy

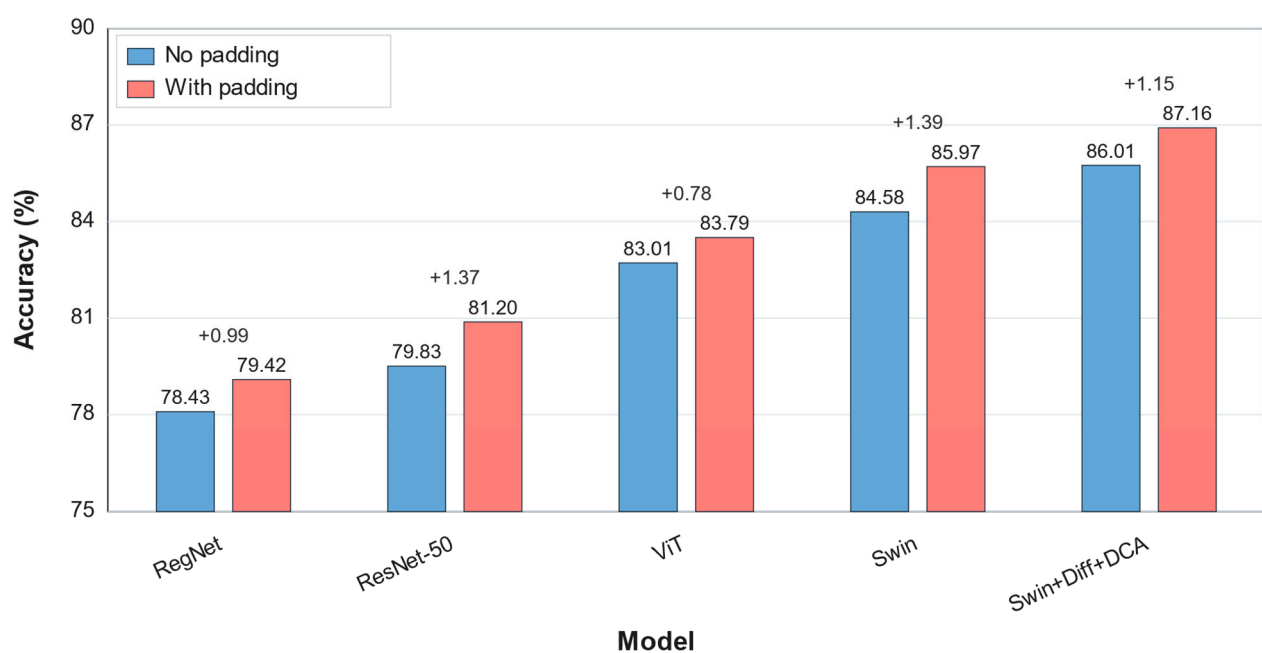


Figure 2. Comparison between direct resizing and aspect-ratio-preserving padding for classification input standardization. The two preprocessing strategies were evaluated under the same data split, model architecture, training settings, and test-set evaluation protocol, and the reported metrics are used to compare their effects on classification performance.

Data augmentation in this paper consists of offline augmentation and online augmentation during training. Offline augmentation is performed during classification dataset construction and is mainly applied to training samples. It includes HSV-space color perturbation, such as changes in hue, saturation, and brightness, to expand the training set and improve model robustness to illumination and color variations. The offline augmented training data are saved in classifier_dataset_hsv. The validation and test sets do not contain randomly augmented samples, avoiding evaluation results being affected by augmentation randomness.

During model training, the training set further uses online random augmentation, including fixed-size resizing, random horizontal flipping, random vertical flipping, random rotation, and ImageNet mean-variance normalization. The validation and test sets use only deterministic preprocessing, including fixed-size resizing, tensor conversion, and ImageNet normalization, without random flipping, random rotation, or color perturbation.

By separating offline and online augmentation, this paper ensures that the training stage contains data perturbation while the final evaluation process remains reproducible and fair.

The augmentation settings were kept consistent within each corresponding stage. During classification dataset construction, offline HSV and brightness augmentation was applied only to the training set, and each training image was augmented with two additional images. The hue shift was randomly sampled from -3 to $+3$ in OpenCV hue units, the saturation factor was sampled from 0.7 to 1.3 , the brightness scaling factor was sampled from 0.7 to 1.3 , and the brightness offset was sampled from -30 to $+30$. During classification model training, the same online augmentation strategy was applied to all classification models, including ResNet, RegNet, ViT, Swin Transformer, and the proposed variants. Specifically, random horizontal flipping and random vertical flipping were applied with probabilities of 0.5 and 0.5 , respectively, and random rotation was performed within -15° to $+15^\circ$. No online hue, saturation, or brightness perturbation was used during classification model training. The validation and test sets did not use random augmentation and were processed only with deterministic resizing, tensor conversion, and ImageNet normalization.

2.3. Model

This study uses Swin Transformer as the base classification network and introduces Differential Attention and Dynamic Channel Aggregation (DCA) into its deep attention and intermediate feature propagation paths. The overall model takes cropped and standardized soybean images as input. It first extracts multi-scale visual features through the hierarchical window attention structure of Swin Transformer. Then, a differential attention branch is introduced in Stage 4 to enhance the model's response to weak inter-class differences. Meanwhile, DCA is introduced in Stage 3 and implemented as a dynamic residual channel routing mechanism to strengthen middle-level local detail reuse. The two modules act on different stages and are designed to jointly improve fine-grained soybean variety classification.

Swin Transformer progressively models image representations through local window self-attention and hierarchical feature downsampling. Compared with traditional convolutional neural networks, this structure can preserve local texture modeling ability while using self-attention to capture longer-range regional relationships. For soybean variety classification, differences among varieties may appear in overall morphology, local texture, and color distribution. Therefore, a Transformer backbone with hierarchical representation ability can better adapt to this fine-grained recognition task.

2.3.1. Dynamic Channel Aggregation

Dynamic Channel Aggregation (DCA) is introduced to alleviate the weakening of middle-level discriminative details during deep feature propagation. Compared with the deep semantic features before final classification, Stage-3 features usually retain more local texture, boundary morphology, and regional color variation information. In this paper, DCA is implemented as a dynamic residual channel routing mechanism from anchor blocks to target blocks, enabling the model to adaptively select and reuse middle-level channel information according to the current sample feature state. This design prevents the classification decision from relying only on final-stage high-level semantic features and preserves more local cues related to variety differences before deep discrimination.

The DCA module is placed inside Stage 3 to enhance effective feature transmission among different blocks. This study selects the 1st, 4th, and 9th blocks in Stage 3 as anchor blocks, and the 11th, 14th, and 17th blocks as target blocks. Given the input feature of a target block, DCA does not replace the original Swin block. Instead, it dynamically generates

fusion weights from different anchor features through a channel gating network [16,17] and injects the weighted routing feature into the target block in a residual form. This design enables the model to adaptively select more valuable middle-level information according to the current sample feature state, rather than using fixed connections among all intermediate features.

Let the output feature of the i -th block in Stage 3 be denoted as X_i . This study selects the outputs of the 1st, 4th, and 9th blocks as anchor features, denoted as A_1 , A_2 , and A_3 , and performs dynamic residual injection at the 11th, 14th, and 17th target blocks. For a target block output T , global average pooling is first performed along the token dimension to obtain a sample-level channel descriptor. Then, a two-layer MLP is used to generate the channel-wise gating weights for different anchors, as shown in Equation (1):

$$G_j = \sigma(W_2 \delta(W_1 \text{GAP}(T))), j = 1, 2, 3 \quad (1)$$

where G_j denotes the channel-wise gating weight of the j -th anchor feature. The routing feature is defined in Equation (2):

$$R = \sum_{j=1}^3 G_j \odot A_j \quad (2)$$

where A_j denotes the output feature of the j -th anchor block, T denotes the current target block feature, G_j denotes the channel-wise gating weight generated from the target feature for the j -th anchor, and R denotes the weighted fused routing feature. Finally, the target block output is updated in a residual form, as shown in Equation (3):

$$T' = T + \gamma_{dca} R \quad (3)$$

where γ_{dca} is a learnable residual strength parameter initialized to 0. This initialization keeps the model behavior close to the original Swin Transformer at the beginning of training, and the model gradually learns whether to introduce middle-level anchor features during subsequent training.

The 1st, 4th, and 9th blocks are selected as anchor blocks to cover middle-level feature states from shallow to deep within Stage 3. The 11th, 14th, and 17th blocks are selected as target blocks to supplement the later part of Stage 3 with texture, edge, and color details retained in earlier blocks. Without changing the number of layers or downsampling path of the original Swin backbone, this setting forms a residual transmission path from middle-level local details to deep discriminative features.

The advantage of DCA is that it strengthens feature reuse within Stage 3 without destroying the pretrained backbone structure. Differences among soybean varieties are often reflected in local texture, color boundaries, and subtle morphological changes, and such information may be more evident in middle-level features. Through anchor-to-target dynamic routing, the model can supplement deeper discriminative representations with local details retained in earlier blocks, thereby alleviating insufficient detail information caused by excessive abstraction in deep representations. Since this module is connected in a residual form and stabilized with zero initialization, it is compatible with pretrained Swin Transformer weights. It should be noted that the anchor-target combination in this paper is an empirical design based on the depth distribution of Stage 3, and its effectiveness is mainly verified through the main ablation experiments. A more fine-grained search of anchor-target combinations can be further studied in future work.

2.3.2. Differential Window Attention

Diff Attention is injected into the Stage-4 deep window attention module of Swin Transformer. This module adds a lightweight differential attention branch outside the original window attention, using differential modeling of attention responses to highlight more discriminative local regions in the current features. Since Stage 4 has a higher-level semantic representation, introducing Diff Attention at this stage can further enhance subtle inter-variety differences near the classification decision, while avoiding premature interference with the formation of shallow texture features.

For the Stage-4 window input feature X , the original Swin window attention first computes the query, key, and value vectors Q , K , and V , and obtains the original attention distribution as shown in Equation (4):

$$A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + B\right) \quad (4)$$

where B is the relative position bias. In addition to the original attention, Diff Attention introduces a lightweight branch and uses lower-dimensional query, key, and value vectors Q_l , K_l , and V_l to compute auxiliary attention, as shown in Equation (5):

$$A_l = \text{Softmax}\left(\frac{Q_l K_l^\top}{\sqrt{d_l}} + B\right) \quad (5)$$

Then, two learnable coefficients are used to construct the differential attention response in Equation (6):

$$A_{\text{diff}} = \lambda_1 A_l - \lambda_2 \text{Detach}(A) \quad (6)$$

where A denotes the original window attention distribution, A_l denotes the lightweight auxiliary attention distribution, and A_{diff} denotes the differential response constructed from the auxiliary attention and the original attention. $\text{Detach}(A)$ is used to block the direct gradient disturbance of the differential branch on the original main attention branch, so that the newly added branch affects the output in a residual enhancement manner. The coefficients λ_1 and λ_2 are obtained from learnable parameters through Sigmoid. The differential response is applied to the lightweight value vector and then projected through a linear layer before being added to the original window attention output in a residual form, as shown in Equation (7):

$$Y = Y_{\text{base}} + \gamma_{\text{diff}} \text{Proj}(A_{\text{diff}} V_l) \quad (7)$$

where λ_{diff} is a learnable residual strength parameter. This module is injected only into Stage 4 because this stage is closer to the classification decision and is suitable for enhancing deep discriminative responses among similar varieties, while avoiding early interference with shallow and middle-level texture features.

From the perspective of interpretability, DCA and Diff Attention correspond to two key levels in the model's discrimination process. DCA acts on Stage 3 of Swin Transformer and mainly focuses on preserving and reusing middle-level local features. Diff Attention acts on Stage 4 and mainly focuses on fine-grained difference enhancement in the deep semantic stage. For soybean variety recognition, differences among varieties are often concentrated in subtle visual cues such as the hilum region, seed coat color, surface texture, edge morphology, and local markings. If the model loses such local information in deep representations or only focuses on common salient regions shared by different varieties, it may easily misclassify visually similar varieties.

Diff Attention is introduced to enhance the response of deep attention to weak difference regions. Ordinary window attention in deep stages may learn common response regions that are useful for classification but insufficiently discriminative. In contrast, fine-grained soybean variety recognition relies more on capturing weak differences among similar categories. This study adds a lightweight differential attention branch to the Stage 4 window attention and highlights more class-discriminative local regions through differential modeling between auxiliary attention responses and original attention responses. Since Stage 4 is closer to the final classification decision, enhancing differential responses at this stage helps improve the model's discrimination ability for similar varieties.

Therefore, DCA and Diff Attention are designed for two complementary problems in fine-grained soybean classification: DCA focuses on whether middle-level local details can be effectively transmitted, while Diff Attention focuses on whether deep attention can highlight inter-class differences. The former provides richer local discriminative cues, and the latter further enhances difference regions before classification decision-making. The subsequent experimental section verifies the above design assumptions through module ablation, ViT-based supplementary experiments, difficult-class analysis, and visualization results. The structure of the DCA module is shown in Figure 3. The differential attention branch introduced into Stage 4 is illustrated in Figure 4.

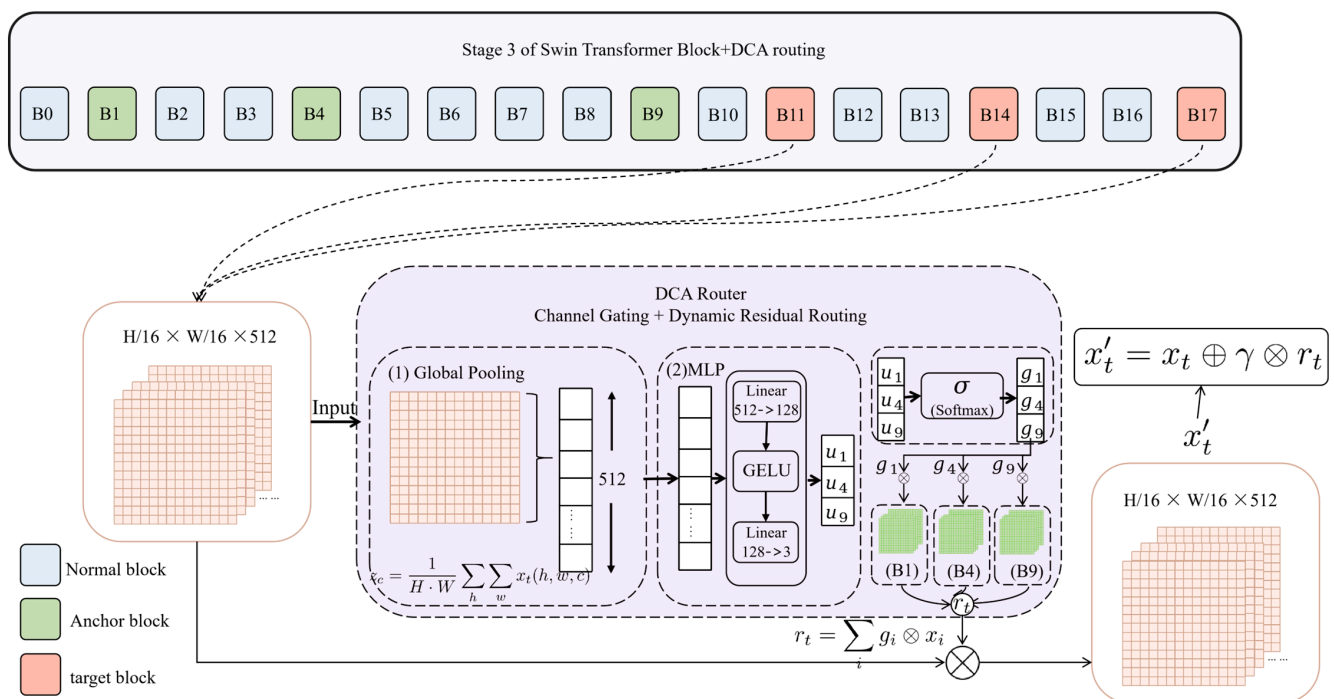


Figure 3. Structure of the proposed DCA module. $\oplus$ denotes addition, and $\otimes$ denotes multiplication

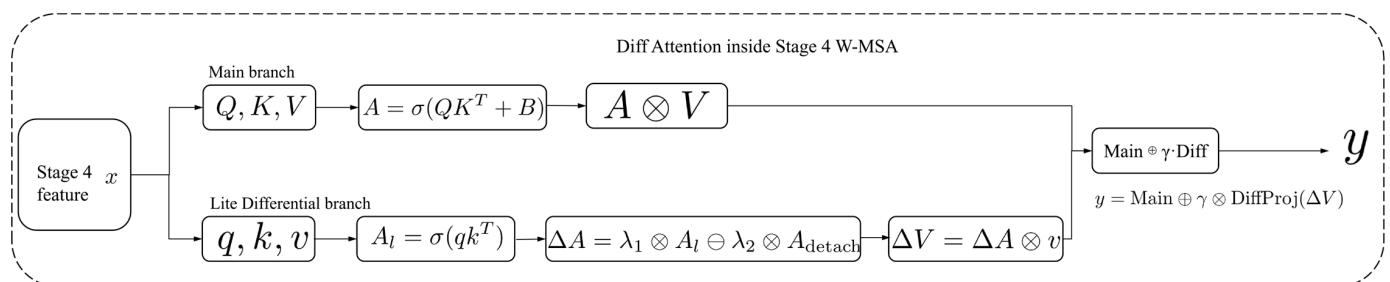


Figure 4. Differential window attention in Stage 4. $\oplus$ denotes addition, $\ominus$ denotes subtraction/difference, and $\otimes$ denotes multiplication

2.4. Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, *F1*-score, Macro-F1, Weighted-F1, and average cross-entropy loss. Accuracy measures the overall classification correctness among all test samples and is defined in Equation (8):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

where *TP*, *TN*, *FP*, and *FN* denote true-positive, true-negative, false-positive, and false-negative samples, respectively.

Precision represents the proportion of samples predicted as a certain class that truly belong to that class, while Recall represents the proportion of samples of a certain class that are correctly recognized. They are defined in Equations (9) and (10):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

F1-score is the harmonic mean of Precision and Recall and is used to comprehensively measure the balance between precision and recall, as shown in Equation (11):

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

Considering the class imbalance in the soybean variety dataset, this paper uses Macro-F1 as the core metric for model selection and comparative analysis. Macro-F1 calculates the equal-weighted average of *F1*-scores across all classes, so it is more sensitive to the recognition performance of minority and difficult categories, as shown in Equation (12):

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (12)$$

where *C* denotes the total number of classes, and *F1_i* denotes the *F1*-score of the *i*-th class. Weighted-F1 is calculated by weighting each class according to its sample size, as shown in Equation (13):

$$\text{Weighted-F1} = \sum_{i=1}^C \frac{n_i}{N} F1_i \quad (13)$$

where *n_i* denotes the number of samples in the *i*-th class, and *n* denotes the total number of samples in the test set. Using Macro-F1 together with Weighted-F1 allows the model's performance on long-tailed categories and overall data distribution to be observed simultaneously.

In the final evaluation stage, all models are tested on an independent test set. The evaluation script outputs Precision, Recall, *F1*-score, support, and average loss for each class and saves overall metrics, per-class metrics, and per-image prediction results. For the final model in this paper, the evaluation script is consistent with the training script in model structure. It first constructs Swin Transformer, then injects Stage-3 DCA and Stage-4 Diff Attention, and finally loads the best_swin_diff_dca_aligned.pth weights for testing, ensuring one-to-one consistency between the training and evaluation structures.

2.5. Implementation Details

During training, ImageNet pretrained weights are used to initialize the Swin Transformer backbone, and label smoothing, grouped learning rates, freeze warm-up, cosine

learning-rate scheduling, mixed-precision training, and gradient clipping are used to improve training stability [18–20].

All model training and experimental evaluation were conducted on a GPU server equipped with an NVIDIA GeForce RTX 5090 GPU with 32 GB memory. The software environment included Python 3.10.0, PyTorch 2.10.0.dev20251203+cu128, Torchvision 0.25.0.dev20251203+cu128, CUDA 12.8, timm 1.0.24, scikit-learn 1.7.2, OpenCV 4.12.0, and Ultralytics 8.3.234. The same GPU server and software environment were used for the main classification experiments, ablation experiments, model complexity analysis, and final test-set evaluation.

All training and evaluation scripts construct data paths, weight paths, and output directories based on the script location, avoiding path errors caused by changes in the running working directory. Evaluation results are uniformly saved in result folders with the same names as the evaluation scripts, including text reports, per-class metrics, overall metric JSON files, and per-image prediction CSV files, facilitating subsequent table generation, confusion analysis, and error sample analysis. To improve experimental reproducibility, the key training configuration, optimization strategies, and module hyperparameters of the main model are summarized in Table 4.

Table 4. Main model training configuration.

Item	Setting
Backbone model	swin_base_patch4_window7_224.ms_in22k_ft_in1k
Input size	224 × 224
Batch size	32
Training epochs	30
Optimizer	AdamW
Backbone learning rate	1×10^{-5}
New modules and classifier learning rate	3×10^{-4}
Weight decay	0.05
Label smoothing	0.05
Freeze warm-up strategy	Freeze the backbone for the first 3 epochs; train only Diff Attention, DCA, and the classifier; then unfreeze all parameters
Diff branch lightweight ratio	LITE_RATIO = 0.25
Diff initialization parameters	INIT_LAMBDA = −2.0, INIT_DIFF_GAMMA = 0.02
DCA anchor blocks	Blocks 1, 4, and 9 in Stage 3
DCA target blocks	Blocks 11, 14, and 17 in Stage 3
DCA residual strength initialization	INIT_DCA_GAMMA = 0.0
Validation/test preprocessing	Deterministic resizing and normalization without random augmentation
Training preprocessing	Offline HSV/color augmentation during dataset construction; online Resize, random horizontal flip, random vertical flip, random rotation, and ImageNet normalization during training.
Hardware platform	GPU server with NVIDIA GeForce RTX 5090 GPU, 32 GB memory
Software environment	Python 3.10.0; PyTorch 2.10.0.dev20251203+cu128; Torchvision 0.25.0.dev20251203+cu128; CUDA 12.8; timm 1.0.24; scikit-learn 1.7.2; OpenCV 4.12.0; Ultralytics 8.3.234

The initialization values of the newly introduced modules were chosen to keep the model close to the pretrained Swin Transformer at the beginning of training. Specifically, γ_{dca} was initialized to 0 so that the DCA path was initially inactive, and γ_{diff} was initialized to a small value to avoid strong perturbation of the pretrained attention output. The lightweight ratio of the Diff branch was set to 0.25 to limit additional computation while retaining sufficient capacity for auxiliary attention modeling.

3. Results

Unless otherwise specified, all classification results reported in this section were obtained on manually verified, cropped, and cleaned soybean seed images rather than on the original whole images. Therefore, the reported metrics evaluate the classification stage of the proposed two-stage framework and should not be interpreted as the end-to-end performance of the complete detection–cropping–classification pipeline.

To improve readability in the result analyses, Dynamic Channel Aggregation (DCA) refers to the Stage-3 residual channel routing module used to reuse middle-level local features, and Differential Attention (Diff Attention) refers to the differential attention branch introduced into the Stage-4 window attention module. In the following tables and discussions, Swin-Diff-DCA denotes the Swin Transformer model equipped with both Diff Attention and DCA.

3.1. Experimental Setup

This section evaluates the soybean variety classification performance of different models on an independent test set. The classification experiments in this paper do not directly perform variety recognition on the original whole images. Instead, they are based on soybean seed images that have been processed by YOLO target localization, manual verification, cropping, and cleaning. Therefore, the results reported in this section mainly reflect the variety recognition ability of classification models on standardized cropped images, rather than the overall performance of a complete end-to-end detection–classification system in complex scenes.

The test set contains 1511 soybean seed images covering 25 soybean varieties. During testing, all input images are resized to 224×224 and processed using a deterministic preprocessing pipeline. Random flipping, random rotation, and color perturbation are not used, ensuring that the evaluation results of different models are comparable and reproducible.

To ensure fair comparison, all classification models use the same training, validation, and test splits and adopt the same deterministic preprocessing and evaluation scripts during validation and testing. To further evaluate the stability of the model comparison, RegNet, ResNet-50, ViT, Swin Transformer, and Swin-Diff-DCA were each trained and evaluated using three different random seeds under the same data split, preprocessing pipeline, and evaluation protocol. The average results across the three runs are reported in Section 3.5. Except for differences in model structure, the training epochs, main optimization strategies, and model selection metric are kept as consistent as possible.

It should be noted that different backbones do not have exactly the same pretrained sources, parameter scales, or structural priors. Therefore, the model comparison in this paper is mainly used to show the performance differences of different methods under the unified experimental protocol, rather than to claim absolute superiority in all scenarios.

3.2. Overall Performance Comparison with Baseline Models

Table 5 shows that Swin-Diff-DCA achieves the best single-run results under the current data split and evaluation protocol.

Table 5. Comparison between the proposed model and baselines.

Model	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro-F1 (%)	Weighted-F1 (%)
RegNet	79.42	71.85	76.28	72.56	79.81
ResNet-50	81.20	75.47	78.32	76.01	81.21
ViT	83.79	77.49	80.95	78.01	84.18
Swin Transformer	85.97	80.26	82.51	80.32	86.21
Swin-Diff-DCA	87.16	81.86	84.65	82.20	87.51

Compared with CNN baseline models, Transformer-based models perform better on this fine-grained classification task. RegNet and ResNet-50 achieve Accuracy values of 79.42% and 81.20%, and Macro-F1 values of 72.56% and 76.01%, respectively. ViT further improves the Accuracy to 83.79% and Macro-F1 to 78.01%. This result suggests that attention-based region relationship modeling is useful for cropped soybean variety classification under the current experimental setting.

Swin Transformer further outperforms ViT, achieving an Accuracy of 85.97% and a Macro-F1 of 80.32%. The 2.18 percentage-point Accuracy gap between Swin Transformer and ViT suggests that hierarchical window attention is better suited to this fine-grained task than directly applying global self-attention to fixed-size cropped inputs. Compared with ordinary ViT, Swin Transformer gradually enhances semantic representation while preserving local texture information through local window attention and hierarchical feature modeling, which is consistent with the local-detail requirement of soybean variety recognition.

After adding Diff Attention and DCA to Swin Transformer, the model performance is further improved. Compared with the Swin baseline, the final model improves Accuracy by 1.19 percentage points, Macro-F1 by 1.88 percentage points, and Weighted-F1 by 1.30 percentage points. Compared with RegNet, the final model improves Accuracy by 7.74 percentage points, Macro-F1 by 9.64 percentage points, and Weighted-F1 by 7.70 percentage points. These results indicate that the proposed module combination improves classification performance under the current test protocol, especially when Macro-F1 is used as the core long-tailed fine-grained evaluation metric.

3.3. Model Complexity Analysis

In addition to classification accuracy, model complexity is also an important factor for evaluating the practical value of soybean variety recognition models. Since the proposed Swin-Diff-DCA model introduces Dynamic Channel Aggregation in Stage 3 and a differential attention branch in Stage 4, it is necessary to examine whether the performance improvement is accompanied by an acceptable computational cost. Therefore, this section compares different models in terms of parameter number, FLOPs, model size, and single-image inference time. The complexity comparison results are shown in Table 6.

Table 6. Model complexity comparison.

Model	Params (M)	FLOPs (G)	Size (MB)	Infer Time (ms)	Macro-F1 (%)	Weighted-F1 (%)
ResNet-50	23.56	8.17	89.87	64.46	76.01	81.21
RegNet	17.96	6.35	68.52	34.58	72.56	79.81
ViT	85.82	35.13	327.37	38.25	78.01	84.18
Swin Transformer	86.77	30.88	331.00	63.80	80.32	86.21
Swin-Diff-DCA	89.66	31.09	342.02	73.64	82.20	87.51

As shown in Table 6, Swin-Diff-DCA introduces a moderate computational overhead compared with the original Swin Transformer baseline. Specifically, the parameter number increases from 86.77 M to 89.66 M, FLOPs increase from 30.88 G to 31.09 G, model size increases from 331.00 MB to 342.02 MB, and single-image inference time increases from 63.80 ms to 73.64 ms under the current testing environment. This additional cost mainly comes from the channel gating and residual routing operations in DCA, as well as the lightweight auxiliary branch used in differential window attention. In return, Swin-Diff-DCA improves Macro-F1 from 80.32% to 82.20% and Weighted-F1 from 86.21% to 87.51% compared with the Swin Transformer baseline. Therefore, the proposed model provides a trade-off between improved fine-grained classification performance and increased inference cost.

From a practical perspective, this trade-off suggests that Swin-Diff-DCA is more suitable for scenarios where recognition accuracy and balanced performance are prioritized over the lowest possible inference latency. For resource-limited or real-time deployment scenarios, the additional inference time should be considered, and further model compression or deployment optimization may be needed. It should also be noted that the current complexity analysis is conducted under the same input size and testing environment, and the reported inference time may vary with hardware platform, software framework, batch size, and deployment settings. Therefore, the results in Table 6 are mainly used for relative comparison among the evaluated models.

3.4. Ablation Experiments on Diff Attention and DCA

Table 7 shows the ablation results of Diff Attention and DCA on the Swin Transformer backbone. Compared with the original Swin Transformer, adding only Diff Attention improves Accuracy from 85.97% to 86.23%, Weighted-F1 from 86.21% to 86.65%, and Macro-F1 from 80.32% to 80.43%. This indicates that introducing a differential attention branch into the Stage 4 deep window attention may help improve discriminative feature representation in the high-level semantic stage under the current experimental setting.

Table 7. Swin ablation experiments. Note: “√” indicates that the corresponding module is used, whereas “-” indicates that it is not used.

Model	Diff Attention	DCA	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
Swin Transformer	-	-	85.97	80.32	86.21
Swin-Diff	√	-	86.23	80.43	86.65
Swin-DCA	-	√	86.30	79.94	86.49
Swin-Diff-DCA	√	√	87.16	82.20	87.51

After adding only DCA, Accuracy increases from 85.97% to 86.30%, and Weighted-F1 increases from 86.21% to 86.49%. However, DCA does not provide a consistent improvement across all evaluation metrics. The Macro-F1 of Swin-DCA decreases from 80.32% to 79.94%, indicating that DCA alone may improve overall prediction correctness but does not stably improve class-balanced performance for minority or difficult varieties. This discrepancy is related to the role of DCA: it strengthens middle-level local feature reuse, but it does not explicitly address class imbalance or highly subtle inter-class differences when used alone. Therefore, DCA should not be interpreted as a standalone solution for difficult-class recognition. Its main value in this study is to provide complementary middle-level local cues, which become more effective when combined with differential attention in the full Swin-Diff-DCA model.

When Diff Attention and DCA are introduced together, the model achieves the best performance. Swin-Diff-DCA reaches 87.16% Accuracy, 82.20% Macro-F1, and 87.51% Weighted-F1, all higher than the models using Diff Attention or DCA alone. In particular, Macro-F1 improves by 1.88 percentage points compared with the Swin baseline, indicating that the combination of the two modules achieves better overall accuracy and Macro-F1 performance on this test set.

From the perspective of model structure, Diff Attention and DCA act at different levels and are therefore complementary. DCA is located in Stage 3 of Swin Transformer and mainly uses anchor block features to assist target blocks, enhancing the transmission of middle-level local texture, color boundary, and morphological details. Diff Attention is located in Stage 4 and mainly enhances differential attention responses in the deep semantic stage close to the classification decision. The former focuses on middle-level local feature reuse, while the latter focuses on deep discriminative feature enhancement. Therefore, combining the two modules leads to more obvious improvements in comprehensive metrics on the current test set.

3.5. Stability Analysis with Multiple Random Seeds

To further evaluate whether the performance improvement is stable under different random initializations, all baseline models and the proposed Swin-Diff-DCA model were trained and evaluated using three different random seeds. As shown in Table 8, Swin-Diff-DCA achieves the highest average Accuracy, Macro-F1, and Weighted-F1 among the compared models. Compared with the Swin Transformer baseline, Swin-Diff-DCA improves the average Accuracy from 84.74% to 87.18%, the average Macro-F1 from 79.08% to 82.33%, and the average Weighted-F1 from 85.02% to 87.44%. These results indicate that the improvement of Swin-Diff-DCA is not only observed in a single run but also remains consistent across different random seeds under the current data split and evaluation protocol.

Table 8. Average results of different models under three random seeds.

Model	Runs	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
RegNet	3	79.15 ± 1.03	72.27 ± 0.66	79.61 ± 0.14
ResNet50	3	79.57 ± 0.81	73.42 ± 1.56	79.96 ± 0.69
ViT	3	83.94 ± 0.33	78.01 ± 0.50	84.21 ± 0.45
Swin Transformer	3	84.74 ± 0.48	79.08 ± 1.09	85.02 ± 0.40
Swin-Diff-DCA	3	87.18 ± 0.61	82.33 ± 0.93	87.44 ± 0.57

3.6. ViT-Based Supplementary Experiment

To further analyze whether the design ideas of Diff Attention and DCA depend only on the hierarchical window structure of Swin Transformer, this study conducts supplementary experiments based on the ViT-B/16 backbone. ViT and Swin Transformer are structurally different: ViT uses global self-attention to model image tokens, while Swin Transformer uses hierarchical window attention to gradually extract multi-scale features. Therefore, transferring Diff Attention and DCA to the ViT backbone can test the applicability of these two modules to fine-grained soybean variety classification from the perspective of a different Transformer structure.

The ViT-based supplementary experiment uses the ViT baseline as the foundation and constructs three variants: ViT + Diff, ViT + DCA, and ViT + Diff + DCA. These models are evaluated on the same test set. The results show that the ViT baseline achieves 83.79% Accuracy, 78.01% Macro-F1, and 84.18% Weighted-F1. ViT + Diff achieves 83.59%, 78.51%, and 83.44%, respectively. ViT + DCA achieves 84.72%, 78.58%, and 84.82%, respectively. ViT + Diff + DCA further improves the metrics to 85.65%, 79.66%, and 85.92%, achieving the best performance among the ViT-series models. The ViT-based supplementary results are summarized in Table 9.

The results show that DCA and Diff + DCA can bring relatively stable performance improvements on the ViT backbone, indicating that the design ideas of middle-level feature reuse and deep differential enhancement do not completely depend on the window attention structure of Swin Transformer. Among them, ViT + DCA improves Accuracy, Macro-F1, and Weighted-F1 compared with the ViT baseline, and ViT + Diff + DCA achieves the best results, suggesting that the two modules also have certain complementarity on the ViT backbone. It should be noted that when Diff Attention is used alone, Macro-F1 increases from 78.01% to 78.51%, but Accuracy and Weighted-F1 decrease to 83.59% and 83.44%, respectively. This indicates that Diff Attention alone mainly improves the class-balanced metric when transferred to ViT, while its improvement in overall performance is not stable. Therefore, the ViT-based supplementary experiment can be regarded as supplementary evidence for the Swin-based main results, but it should not replace the main ablation conclusion of Swin-Diff-DCA.

Table 9. ViT-based supplementary experiment. Note: “√” indicates that the corresponding module is used, whereas “-” indicates that it is not used.

Model	Diff Attention	DCA	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
ViT	-	-	83.79	78.01	84.18
ViT + Diff	√	-	83.59	78.51	83.44
ViT + DCA	-	√	84.72	78.58	84.82
ViT + Diff + DCA	√	√	85.65	79.66	85.92

3.7. Confusion Matrix Analysis

The confusion matrix can directly reflect the misclassification relationships among different classes. Overall, the confusion matrix of the final model shows a clear diagonal distribution, indicating that most classes can be correctly recognized. This observation is consistent with the model’s Accuracy of 87.16% and Weighted-F1 of 87.51%, suggesting that the model has good overall classification ability.

However, the off-diagonal elements in the confusion matrix show that the errors are concentrated among a small number of visually similar or class-boundary-ambiguous varieties rather than being uniformly distributed across all classes. For example, 17 samples of lk314 are misclassified as zld105, 12 samples of hd16 are misclassified as jng20839, and 9 samples of zd57 are misclassified as zd53. In addition, 5 and 6 samples of nn55 are misclassified as zh301 and xzd1, respectively, and 4 samples of zd61 are misclassified as zd53.

The confusion among zd-series varieties is particularly noteworthy. The misclassifications from zd57 to zd53 and from zd61 to zd53 indicate that the model still has difficulty distinguishing these visually similar varieties. These classes may have high similarity in seed morphology, seed coat color, surface texture, and local markings, making it difficult for the model to distinguish them stably using only a single cropped image. This result is also consistent with the difficult-class analysis above, indicating that the errors are mainly concentrated among fine-grained visual neighbor classes rather than being randomly distributed across all classes.

To further quantify the recognition performance of difficult classes, Table 10 presents the per-class metrics of Swin-Diff-DCA on several classes with low F1-scores or high confusion.

Table 10. Metrics of Swin-Diff-DCA on highly confused classes.

Class	Precision (%)	Recall (%)	F1 (%)	Support	Major Misclassification Direction of True Class
zd51	33.33	58.33	42.42	12	zd51→sd29:2; zd51→nn55/sd30/xzd1: 1 each
zd59	42.86	54.55	48.00	11	zd59→nn55/sn23/xd18/zd51/zd53: 1 each
jng20839	48.48	88.89	62.75	18	jng20839→nn43: 2
zd53	53.49	82.14	64.79	28	zd53→zd51: 2
zd57	86.84	57.89	69.47	57	zd57→zd53: 9; zd57→nn55/zd51: 3 each; zd57→xd18/zd61: 2 each

From Table 10, zd51 and zd59 are the most difficult low-sample classes in the current test set. They contain only 12 and 11 test samples, respectively, and their F1-scores are 42.42% and 48.00%. These low F1-scores indicate that the final model still has limited reliability for some minority varieties. For such low-sample classes, even a small number of wrong predictions can cause large fluctuations in Precision, Recall, and F1-score. In addition, the misclassification directions show that these errors are not random. For example, zd51 is confused with sd29, nn55, sd30, and xzd1, while zd59 is confused with nn55, sn23, xd18, zd51, and zd53. This suggests that low sample size and visual similarity jointly contribute to the weak performance of these classes.

The confusion matrix shown in Figure 5 can directly reflect the classification performance of the model for each soybean variety. The confusion matrix analysis also shows that several errors occur among visually similar soybean varieties, especially within the zd-series classes. For example, zd57 is frequently misclassified as zd53, and zd53 also shows confusion with zd51. These varieties may share similar seed coat color, contour, texture, and local markings, making them difficult to distinguish using only cropped RGB images. Therefore, although Swin-Diff-DCA improves the overall classification performance, its practical reliability is still limited for low-sample and visually similar varieties. Future work should increase samples for difficult classes such as zd51 and zd59 and further explore class-balanced learning, hard example mining, metric learning, or local-region modeling to improve recognition of these challenging categories.

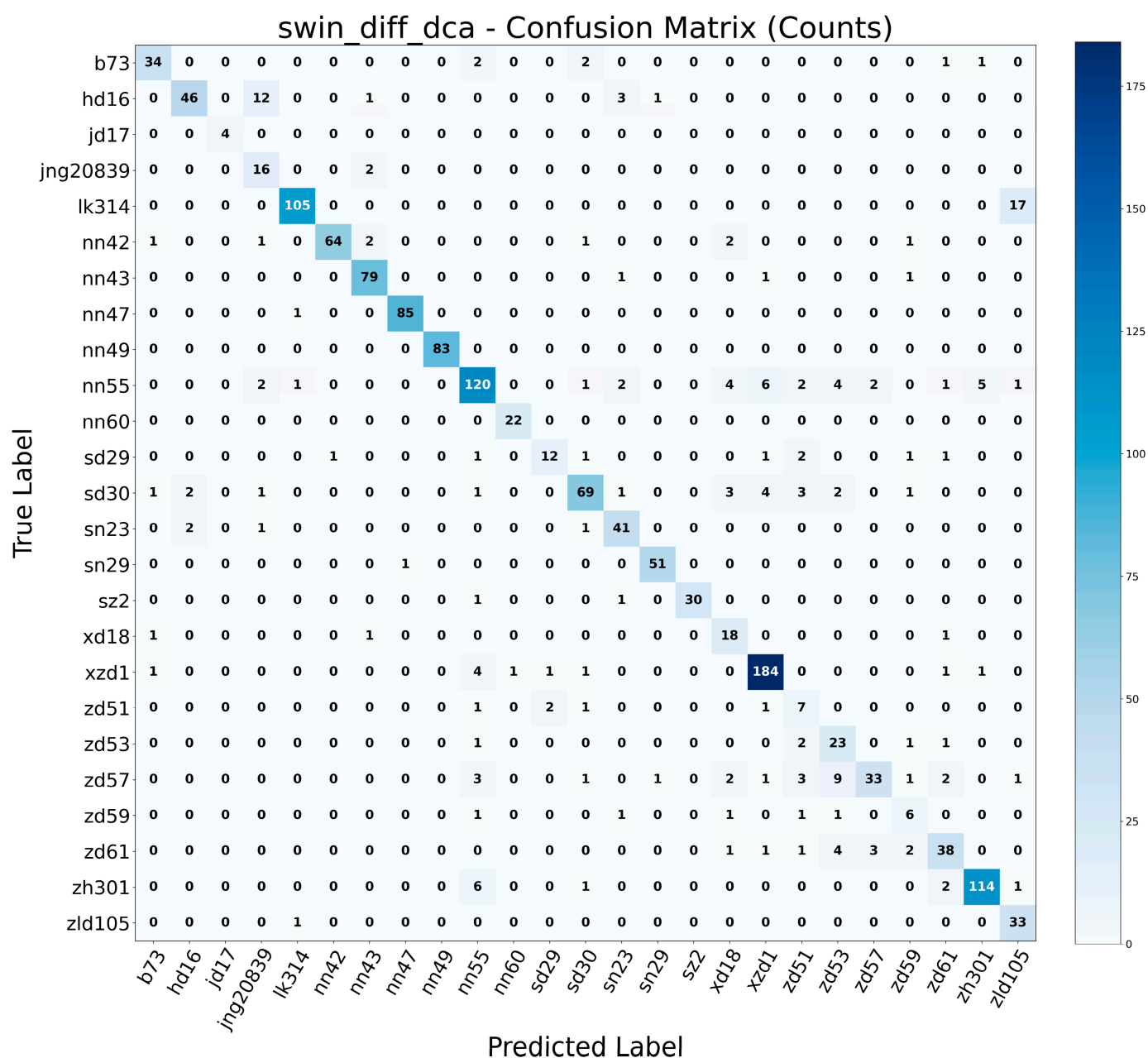


Figure 5. Confusion matrix of the Swin-Diff-DCA model on the soybean variety test set.

4. Discussion

4.1. Difference Between Macro-F1 and Weighted-F1

The difference between Macro-F1 and Weighted-F1 can reflect the influence of class imbalance on model evaluation. The final model achieves an Accuracy of 87.16% and a Weighted-F1 of 87.51%, but its Macro-F1 is 82.20%. Since Weighted-F1 weights classes according to the number of samples, large-sample classes have a greater influence on this metric. In contrast, Macro-F1 treats all classes equally and is therefore more affected by the performance of minority and difficult classes.

This metric gap indicates that reporting only Accuracy may overestimate the model's balanced recognition ability across all varieties. The final model performs well on the majority of samples, resulting in high Accuracy and Weighted-F1. However, the lower F1-scores of classes such as zd51, zd59, jng20839, and zd53 reduce the Macro-F1. This shows that although the model has strong overall performance, it still has limitations in recognizing minority and easily confused classes.

Therefore, it is necessary to report Accuracy, Macro-F1, and Weighted-F1 together. Accuracy reflects the overall correctness, Weighted-F1 reflects the sample-distribution-weighted comprehensive performance, and Macro-F1 focuses more on whether each class can be recognized fairly. For fine-grained soybean variety classification, Macro-F1 is especially important because errors on minority and similar varieties can also affect the reliability of variety verification and germplasm resource management in real applications.

4.2. Discussion on Model Effectiveness

The experimental results are generally consistent with the design motivation of the proposed model. Soybean variety classification relies on local texture, seed coat color, edge morphology, and subtle structural differences, and these features appear differently at different network levels. Shallow and middle-level features usually preserve more local texture and morphological information, while deep features are closer to semantic classification decisions. Therefore, relying only on an ordinary classification backbone may make it difficult to simultaneously consider middle-level details and high-level discriminative differences.

Swin Transformer outperforms CNN baselines mainly because its hierarchical window attention structure is more suitable for modeling local and contextual relationships in soybean images. CNNs have strong local texture modeling ability, but their ability to explicitly model relationships among different regions is relatively limited. Swin Transformer uses window attention and hierarchical downsampling to preserve local structural representation while gradually forming high-level semantic features, thereby better capturing subtle and distributed visual differences among soybean varieties.

The effectiveness of Diff Attention is mainly reflected in deep discriminative feature enhancement. This module is introduced into Stage 4, which is closer to the final classification decision and has stronger semantic abstraction ability. Through the differential attention branch, the model can enhance class-discriminative responses in deep window attention, helping it distinguish visually similar varieties. In the ablation experiment, Swin-Diff improves Accuracy and Weighted-F1 compared with the Swin baseline, indicating that Diff Attention alone brings limited improvements. Swin-Diff shows only small single-run gains in Accuracy and Weighted-F1, so the effect of Diff Attention alone should be interpreted cautiously. However, the improvement is small, and its stability should be judged together with the results of the combined model.

The effectiveness of DCA is mainly reflected in middle-level feature reuse. This module is introduced into Stage 3 and uses dynamic residual channel routing from anchor blocks to target blocks, enabling the model to adaptively select useful channels according to the current sample. Since Stage-3 features still contain rich texture, boundary, and color variation information, DCA helps alleviate the weakening of these local details during deep abstraction. Although Accuracy and Weighted-F1 improve when DCA is used alone, its Macro-F1 slightly decreases, indicating that DCA alone is more helpful for overall prediction than for class-balanced recognition of minority or difficult categories. Therefore, the effect of DCA should be interpreted together with differential attention rather than as an independent solution to long-tailed fine-grained classification.

The combination of Diff Attention and DCA achieves the best results on the current test set, supporting the judgment that the two modules have certain complementarity in this task. DCA mainly enhances middle-level local feature transmission, while Diff Attention mainly enhances high-level differential attention responses. The former provides richer local discriminative cues, and the latter strengthens inter-class differences before classification decision-making. By combining the two modules, the model can use both middle-level details and deep semantic discriminative information, achieving the best Accuracy, Macro-F1, and Weighted-F1.

4.3. Interpretability Analysis

To further analyze the model's decision basis, this paper uses Grad-CAM [21] to compare the visualization results of the Swin Transformer baseline and Swin-Diff-DCA. The visualized samples include correctly classified samples and misclassified samples. By comparing the heatmap regions of the two models, this analysis observes whether the proposed model pays more attention to discriminative regions related to soybean seed texture, edge morphology, and local color differences.

From Figure 6, it can be observed that Swin-Diff-DCA places more heatmap responses on the seed body, edge, texture, and color variation regions in some correctly classified samples, while the responses of the Swin baseline are relatively more scattered or background-influenced in some cases. This observation suggests that the proposed modules may help enhance the model's ability to represent fine-grained discriminative features. However, Grad-CAM is mainly a qualitative visualization tool, and the heatmap results should be regarded as auxiliary evidence rather than direct proof of the internal mechanism of the model.

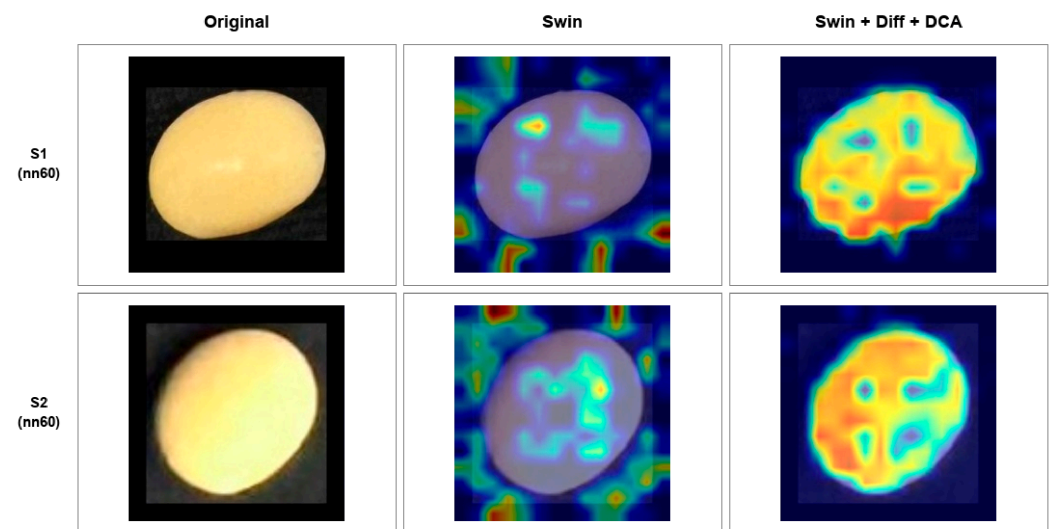


Figure 6. Grad-CAM visualization comparison between Swin Transformer and Swin-Diff-DCA. The middle panels show representative correctly classified cases, while the right panels show representative incorrectly classified cases.

4.4. Limitations

Although the proposed Swin-Diff-DCA model achieves the best performance on the current test set, it still has certain limitations. First, the test set contains 1511 images covering 25 classes, which is still limited for comprehensively evaluating model robustness. Some classes have very few test samples. For example, jd17 has only 4 samples, zd59 has 11 samples, and zd51 has 12 samples. For these low-sample classes, a small number of prediction errors can cause obvious fluctuations in Precision, Recall, and F1-score. Therefore, the corresponding metrics should not be overinterpreted.

Second, the current results should be interpreted in relation to real-world soybean seed inspection scenarios. The proposed two-stage framework is applicable to scenarios such as laboratory seed variety verification, germplasm sample archiving, and batch-level seed quality inspection, where images can be acquired under relatively controlled conditions and candidate seed regions can be verified or cleaned before classification. In such scenarios, YOLO-based localization and cropped-image classification can reduce background interference and support efficient screening. However, the present evaluation is still limited because it is based on manually verified, cropped, and cleaned soybean seed images rather

than original whole images in an uncontrolled deployment pipeline. In practical use, detection-box quality, incomplete cropping, overlapping seeds, image blur, illumination changes, background variation, device differences, and unknown or low-sample varieties may affect the final recognition results. Therefore, the reported classification performance should be regarded as evidence for the cropped-image classification stage, and future deployment requires independent evaluation of the complete YOLO detection, cropping, and classification pipeline under real inspection conditions.

Third, the current experiments mainly demonstrate the effectiveness of the model under this data split, but its generalization ability across years, locations, devices, and acquisition environments has not yet been verified. The appearance of soybean seeds may be affected by growth environment, acquisition device, shooting conditions, and sample processing methods. Therefore, validation on external independent datasets is needed in future work to further support the application value of the model in real agricultural scenarios.

Future work should first expand data for low-sample classes, especially zd51, zd59, and jd17, and then evaluate class-balanced loss, focal loss [22], resampling, or hard example mining as separate experimental variables. Metric learning [23,24] or contrastive learning can also be explored to enlarge feature distances among visually similar varieties such as the zd-series classes. In addition, more interpretability analysis can be used to verify whether the model consistently focuses on effective discriminative regions such as the hilum, texture, edge morphology, and color variation.

Although this paper reports the parameter number, FLOPs, model size, and single-image inference time of the evaluated models, the current efficiency analysis is still limited to the classification stage under a fixed testing environment. The reported inference time may vary across hardware platforms, software frameworks, batch sizes, and deployment settings. Therefore, future work should further evaluate the complete YOLO detection, cropping, and classification pipeline on target deployment devices, so as to better assess the practical efficiency of the whole soybean variety recognition system.

5. Conclusions

This study constructs a cascaded soybean variety recognition process consisting of YOLO-based object localization, manual verification and cropping, and fine-grained variety classification. The quantitative results reported in this paper mainly reflect the performance of the classification model on manually verified, cropped, and cleaned soybean seed images, rather than the deployment-level performance of a fully end-to-end recognition system on original whole images. This design explicitly separates target localization, sample quality control, and cropped-image classification, allowing the classification model to focus on soybean seed regions rather than background, scale variation, or irrelevant image content.

To improve cropped soybean variety classification, this paper proposes Swin-Diff-DCA based on Swin Transformer. Dynamic Channel Aggregation (DCA) is introduced in Stage 3 and implemented as a dynamic residual channel routing mechanism to reuse middle-level texture, color-boundary, contour, and local structural information. Diff Attention is injected into the Stage-4 window attention to strengthen deep differential response modeling for visually similar varieties. These two modules act on middle-level feature reuse and deep discriminative enhancement, respectively, forming a complementary structure for fine-grained soybean classification.

Under the current data split and evaluation protocol, Swin-Diff-DCA achieves average results of 87.18% Accuracy, 82.33% Macro-F1, and 87.44% Weighted-F1 across three random seeds on 25 soybean varieties and 1511 test images, outperforming RegNet, ResNet-50, ViT, the Swin Transformer baseline, and the single-module Swin variants. The ablation results show that Diff Attention and DCA provide complementary benefits when combined, while

the confusion matrix and per-class metrics indicate that low-sample and visually similar classes such as zd51, zd59, jng20839, zd53, and zd57 remain the main sources of errors.

Future work should expand the number and diversity of samples for difficult classes, evaluate class-imbalance strategies as separate experimental variables, and verify generalization across different years, devices, and acquisition environments [25]. In addition, future work should further evaluate the complete end-to-end detection–classification pipeline, including detection quality, cropping reliability, and inference efficiency on target deployment devices, before making deployment-oriented claims.

Author Contributions: Conceptualization, P.S. and J.H. (Ji Huang); methodology, P.S.; software, P.S.; validation, P.S.; formal analysis, P.S.; investigation, P.S., Q.Y., Z.C. and L.Z.; resources, H.X. and J.H. (Ji Huang); data curation, P.S. and Q.Y.; writing—original draft preparation, P.S.; writing—review and editing, J.H. (Ji Huang) and J.H. (Junxian Huang); visualization, P.S.; supervision, J.H. (Ji Huang) and J.H. (Junxian Huang); project administration, J.H. (Junxian Huang). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement: Ethical review and approval were not required for this study because it did not involve human participants, human data, animal subjects, or animal-derived materials.

Data Availability Statement: The code generated during the current study is publicly available in the GitHub repository soybean_identify at https://github.com/a-little-programmer/soybean_identify (accessed on 10 June 2026). The datasets generated and analyzed during the current study are not publicly available due to ongoing use of the variety image resources, but are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
CNN	Convolutional neural network
DCA	Dynamic Channel Aggregation
Diff Attention	Differential attention
FLOPs	Floating-point operations
Grad-CAM	Gradient-weighted Class Activation Mapping
HSV	Hue, saturation, and value
ViT	Vision Transformer
YOLO	You Only Look Once

References

1. Jin, C.; Zhou, L.; Pu, Y.; Zhang, C.; Qi, H.; Zhao, Y. Application of deep learning for high-throughput phenotyping of seed: A review. *Artif. Intell. Rev.* **2025**, *58*, 76. [\[CrossRef\]](#)
2. Zhu, S.; Zhou, L.; Zhang, C.; Bao, Y.; Wu, B.; Chu, H.; Yu, Y.; He, Y.; Feng, L. Identification of Soybean Varieties Using Hyperspectral Imaging Coupled with Convolutional Neural Network. *Sensors* **2019**, *19*, 4065. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Zhu, S.; Zhang, J.; Chao, M.; Xu, X.; Song, P.; Zhang, J.; Huang, Z. A Rapid and Highly Efficient Method for the Identification of Soybean Seed Varieties: Hyperspectral Images Combined with Transfer Learning. *Molecules* **2020**, *25*, 152. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Wei, Y.; Li, X.; Pan, X.; Li, L. Nondestructive Classification of Soybean Seed Varieties by Hyperspectral Imaging and Ensemble Machine Learning Algorithms. *Sensors* **2020**, *20*, 6980. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Zhang, L.; Sun, L.; Jin, X.; Zhao, X.; Li, S. DAFFnet: Seed classification of soybean variety based on dual attention feature fusion networks. *Crop J.* **2025**, *13*, 619–629. [\[CrossRef\]](#)
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [\[CrossRef\]](#)

7. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [[CrossRef](#)]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
9. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollar, P. Designing Network Design Spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10428–10436. [[CrossRef](#)]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [[CrossRef](#)]
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021. [[CrossRef](#)]
12. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [[CrossRef](#)]
13. Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN Models for Fine-Grained Visual Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457. [[CrossRef](#)]
14. Fu, J.; Zheng, H.; Mei, T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4476–4484. [[CrossRef](#)]
15. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [[CrossRef](#)]
16. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)] [[PubMed](#)]
17. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision-ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 3–19. [[CrossRef](#)]
18. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [[CrossRef](#)]
19. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017. [[CrossRef](#)]
20. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019. [[CrossRef](#)]
21. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [[CrossRef](#)]
22. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007. [[CrossRef](#)]
23. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823. [[CrossRef](#)]
24. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In Proceedings of the Advances in Neural Information Processing Systems 33, Virtual, 6–12 December 2020; pp. 18661–18673. [[CrossRef](#)]
25. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35. Available online: <https://www.jmlr.org/papers/v17/15-239.html> (accessed on 30 May 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.