agriculture

MDPI

Article

U-MoEMamba: A Hybrid Expert Segmentation Model for Cabbage Heads in Complex UAV Low-Altitude Remote Sensing Scenarios

Rui Li [1], Xue Ding [2,3,4,*], Shuangyun Peng [2,3,4] and Fapeng Cai [2]

[1] School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China; 2424410008@ynnu.edu.cn
[2] Faculty of Geography, Yunnan Normal University, Kunming 650500, China; pengshuangyun@ynnu.edu.cn (S.P.); 234130115@ynnu.edu.cn (F.C.)
[3] Key Laboratory of Resources and Environmental Remote Sensing for Universities in Yunnan, Kunming 650500, China
[4] Center for Geospatial Information Engineering and Technology of Yunnan Province, Kunming 650500, China
* Correspondence: 4228@ynnu.edu.cn

Abstract

To address the challenges of missed and incorrect segmentation in cabbage head detection under complex field conditions using UAV-based low-altitude remote sensing, this study proposes U-MoEMamba, an innovative dynamic state-space framework with a mixture-of-experts (MoE) collaborative segmentation network. The network constructs a dynamic multi-scale expert architecture, integrating three expert paradigms—multi-scale convolution, attention mechanisms, and Mamba pathways—for efficient and accurate segmentation. First, we design the MambaMoEFusion module, a collaborative expert fusion block that employs a lightweight gating network to dynamically integrate outputs from different experts, enabling adaptive selection and optimal feature aggregation. Second, we propose an MSCrossDualAttention module as an attention expert branch, leveraging a dual-path interactive attention mechanism to jointly extract shallow details and deep semantic information, effectively capturing the contextual features of cabbages. Third, the VSSBlock is incorporated as an expert pathway to model long-range dependencies via visual state-space representation. Evaluation on datasets of different cabbage growth stages shows that U-MoEMamba achieves an mIoU of 89.51% on the early-heading dataset, outperforming SegMamba and EfficientPyramidMamba by 3.91% and 1.4%, respectively. On the compact heading dataset, it reaches 91.88%, with improvements of 2.41% and 1.65%. This study provides a novel paradigm for intelligent monitoring of open-field crops.

Keywords: UAV-based low-altitude remote sensing; cabbage head segmentation; mixture of experts (MoE); attention mechanism; U-MoEMamba

check for updates

Academic Editor: Andrzej Borusiewicz

Received: 9 July 2025
Revised: 7 August 2025
Accepted: 7 August 2025
Published: 9 August 2025

Citation: Li, R.; Ding, X.; Peng, S.; Cai, F. U-MoEMamba: A Hybrid Expert Segmentation Model for Cabbage Heads in Complex UAV Low-Altitude Remote Sensing Scenarios. *Agriculture* **2025**, *15*, 1723. https://doi.org/10.3390/agriculture15161723

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cabbage is widely cultivated worldwide due to its ease of cultivation, rich nutritional value, and crisp, juicy texture. As the fourth most important vegetable crop worldwide [1], the phenotypic traits of cabbage heads—such as size and compactness—directly determine their commercial value. Traditionally, monitoring cabbage growth has primarily relied on manual surveys [2], which are time-consuming, labor-intensive, and prone to errors induced by subjectivity. In high-altitude regions, the cost of monitoring is significantly higher than in the plains. Moreover, the spatial resolution of most high-resolution satellite

imagery remains at the meter level, insufficient to meet the accuracy demands of precision agriculture at the individual plant scale [3]. The advent of unmanned aerial vehicle (UAV)-based low-altitude remote sensing offers centimeter-level spatial resolution, providing a novel solution for building fine-grained and digitized monitoring systems for cabbages. Particularly in complex terrains such as the Yunnan Plateau, UAV-based approaches can be 15–30 times more efficient than traditional methods [4].

With the rapid advancement of computer vision, semantic segmentation techniques have evolved from traditional machine learning methods to deep learning paradigms. In precision agriculture, Huang et al. [5] utilized fully convolutional networks (FCNs) by replacing fully connected layers with convolutional layers to achieve pixel-level weed–rice segmentation. Sahin et al. [6] employed a U-Net with a symmetric encoder–decoder architecture and skip connections to effectively fuse semantic and spatial featu signifyres for crop–weed discrimination. Kang et al. [7] enhanced DeepLabV3+ with attention mechanisms to assess the maturity of broccoli. Crespo et al. [8] introduced the Mask R-CNN instance segmentation model for efficient strawberry segmentation. These innovations mark a new era of pixel-level intelligent perception in agricultural image analysis, enabling plant-level monitoring and supporting the transition from traditional to digital farming.

Despite progress, existing segmentation methods still struggle in complex farmland environments. Most rely on convolutional neural networks (CNNs), which are inherently limited by their local receptive fields and perform poorly in modeling long-range dependencies, leading to lower segmentation accuracy under occlusion, background clutter, and varied lighting [9,10]. To address the inherent limitations of convolutional neural networks (CNNs), segmentation networks have evolved through several key technological advances. One major development is Transformer-based segmentation [11,12]. For instance, Xiang et al. [13] introduced MixSegNext, which utilizes multi-scale spatial–channel attention to accurately identify chili and pepper clusters. Cai et al. [14] embedded attention modules into PSPNet to enhance weed detection precision in pineapple fields. Similarly, Liu et al. [15] proposed a segmentation method called AISOA-SSformer based on the Transformer architecture for identifying rice leaf diseases, achieving an MIoU of 83.1%, a Dice coefficient of 80.3%, and a recall of 76.5%. Another significant advancement is the integration of hybrid CNN-Transformer models, combining the strengths of both architectures for improved segmentation. Meng et al. [16] combined Shifted Window Transformers with RCNN to enable fast pineapple fruit detection. Wang et al. [17] developed DualSeg, which merges CNN and Transformer modules to identify grape clusters in complex vineyard environments, achieving an F1 score of 0.91 at 83.7 frames per second. Gou et al. [18] proposed CTFFNet, a hybrid CNN-Transformer network for rice field weed segmentation, boosting accuracy to 72.8% in mean IoU. Furthermore, Jeon et al. [19] utilized a similar hybrid architecture to classify wheat varieties and growth stages, reaching an accuracy of 94.05%.

More recently, the Mamba architecture has emerged as a novel approach to visual segmentation, offering efficient training and inference grounded in state-space models (SSM) [20] and hardware-optimized design. Vision Mamba (ViM) [21] and VMamba [22] extend SSM principles to vision tasks, with ViM employing bidirectional SSMs and positional embeddings to model global context and enable position-aware recognition, demonstrating robust performance in dense prediction tasks. Within agricultural applications, Mamba-based models have shown considerable promise. For example, Shi et al. [23] integrated Vision Mamba with ConvNeXt in a UNet framework for quantifying tomato wilt disease. Zhao et al. [24] applied Mamba-based vision systems on UAVs to assess blueberry maturity. Additionally, Zhang et al. [25] introduced GMamba for grape leaf disease segmentation, effectively modeling global context while reducing the computational overhead typical of Transformer-based methods.

Current research on cabbage segmentation focuses primarily on distinguishing cabbage plants from surrounding weeds in complex field environments, as this task plays a critical role in supporting automated weeding, yield estimation, and overall field management in precision agriculture. For instance, Gao et al. [26] employed ConvNeXt as the backbone, integrated the RepVgg structure, and applied the Sigmoid linear unit activation function based on the DeepLabV3+ model, effectively distinguishing between cabbage heads and weeds. Kong et al. [27] proposed a lightweight segmentation network, LCSNet, for cabbage–weed discrimination. Ma et al. [28] introduced an improved U-Net combining multi-scale inputs and attention mechanisms, achieving an mIoU of 88.96%. Tian et al. [29] enhanced YOLOv8n-seg for accurate cabbage head segmentation under complex conditions.

Accurate segmentation of edible cabbage heads is vital for crop monitoring, growth assessment, maturity estimation, and precision cultivation. However, this task remains technically challenging due to the structural similarity of cabbage heads and the complexity of field conditions—such as varying illumination, irregular morphology, overlapping leaves, and similar textures. To achieve precise segmentation, models must possess high spatial resolution, strong feature extraction capabilities, and the ability to model long-range dependencies. Despite recent advances, existing methods often suffer from under-segmentation and mis-segmentation in real-world scenarios involving extreme lighting, dense occlusion, and subtle texture variations. These challenges expose their limitations in fine-grained feature extraction and contextual modeling. Most current approaches rely either on convolutional operations with limited receptive fields or on attention mechanisms that are computationally expensive and insufficient for capturing fine local structures. Few methods can effectively integrate both global context and local detail under such complex conditions.

To address this gap, we are inspired by recent advances in spectral clustering techniques that combine K-means and GMM [30]. Building on this idea, we propose U-MoEMamba, a hybrid expert collaborative segmentation network built upon a dynamic and differentiable state space framework. This model leverages Mamba's strength in capturing long-range dependencies while enhancing local feature extraction, resulting in significant performance improvements. The main contributions of this work are as follows:

(1). Design of a Heterogeneous Expert Collaboration Framework: This framework features a tri-expert architecture that integrates multi-scale convolutional experts, cross-attention experts, and Mamba path experts. The fusion leverages the strengths of convolutional neural networks (CNNs) for local perception, attention mechanisms for global context modeling, and state-space models (SSMs) for capturing long-range dependencies.

(2). Proposal of the MambaMoEFusion Module: We construct a hybrid MoE fusion module, which includes three distinct expert pathways and a lightweight gating network for adaptive integration. Unlike traditional fusion strategies such as concatenation or static weighting, this dynamic approach allows context-aware selection of the most suitable expert features.

(3). Development of the MSCrossDualAttention Module: We propose a dual-path attention mechanism that simultaneously processes features along the spatial and channel dimensions, aligning shallow structural details with deep semantic information and thereby enhancing feature complementarity.

(4). Lightweight Gating via MOEGatingNetwork: We design a resource-efficient gating mechanism. By generating expert weights through adaptive average pooling and fully connected layers, our approach achieves hardware-friendly dynamic resource allocation with minimal computational overhead.

The proposed U-MoEMamba model advances the innovative application of dynamic expert collaboration and long-range dependency modeling in the intelligent analysis of UAV-based crop imagery. Theoretically, the heterogeneous expert fusion framework integrates multi-scale convolution, cross-attention mechanisms, and Mamba-based state-space modeling, offering a novel architectural paradigm for high-precision segmentation of fine-grained plant structures. Practically, experimental results on cabbage datasets across different growth stages demonstrate the model's ability to accurately segment edible cabbage heads under challenging conditions such as leaf occlusion and varying illumination. The findings provide a transferable reference framework for intelligent crop monitoring, growth assessment, and decision-making in precision agriculture.

2. Materials and Methods

2.1. Data Acquisition

The empirical samples in this study consist of heading cabbage. Data collection was conducted from 1–3 May 2025, in Tonghai County, Yuxi City, Yunnan Province, China—one of the country's major open-field vegetable production regions. The specific sampling area was located in Dashu Community, Xiushan Subdistrict, with geographic coordinates ranging from 102°30′25″ E to 102°52′53″ E and 23°55′11″ N to 24°14′49″ N. The region is characterized by a subtropical plateau monsoon climate, with an average annual temperature of approximately 15.8 °C and an average annual precipitation of about 900 mm. The geographic location of the study area is shown in Figure 1a. Image data were collected using a DJI Phantom 4 UAV platform (Figure 1b), manufactured by DJI Innovations, Shenzhen, China. The UAV was equipped with a high-resolution RGB camera, a centimeter-level Real-Time Kinematic (RTK) positioning system, and a 3-axis gimbal stabilizer. UAV operations were performed in a nadir-view configuration, with a flight altitude of approximately 3 m above ground level (relative to the crop canopy).

The dataset spans the full growth cycle of cabbage, with a particular focus on two critical phenological stages: the early heading stage and the compact heading stage (see Figure 1d,e). To ensure the diversity and representativeness of the data, images were acquired from multiple field plots under varying environmental conditions. These included different lighting scenarios (overcast/sunny), planting densities, and levels of leaf occlusion, as illustrated in Figure 2. After data screening and quality control, a total of 1900 valid image samples were obtained, including 1200 images from the compact heading stage and 700 from the early heading stage. All images were captured at a resolution of 5472 × 3648 pixels. Manual annotation was performed using the open-source tool LabelMe. For each image, annotators manually delineated key points along the contours of target objects and connected them into smooth, closed polygons that closely match the actual shape of the cabbage heads. The annotated targets mainly include two key structural components: the cabbage head (leaf ball) and the outer leaves. All annotations were exported in JSON format, providing structured label data to support downstream training and evaluation. An example of the annotation process is shown in Figure 3.

2.2. Data Augmentation

To enhance the generalization ability of the model and mitigate the risk of overfitting, a range of data augmentation techniques was applied to the cabbage image dataset. These augmentations aimed to increase sample diversity and improve the robustness of the segmentation model under varying field conditions. Specifically, the following augmentation strategies were employed:

(5). Horizontal Flipping: Images were randomly flipped horizontally with a probability of 50%.

(6). Saturation Adjustment: Color saturation was linearly perturbed within the range of −25% to +25% to simulate varying lighting and imaging conditions.

(7). Salt-and-Pepper Noise Injection: Random salt-and-pepper noise was introduced to 0.1% of image pixels to enhance noise robustness.

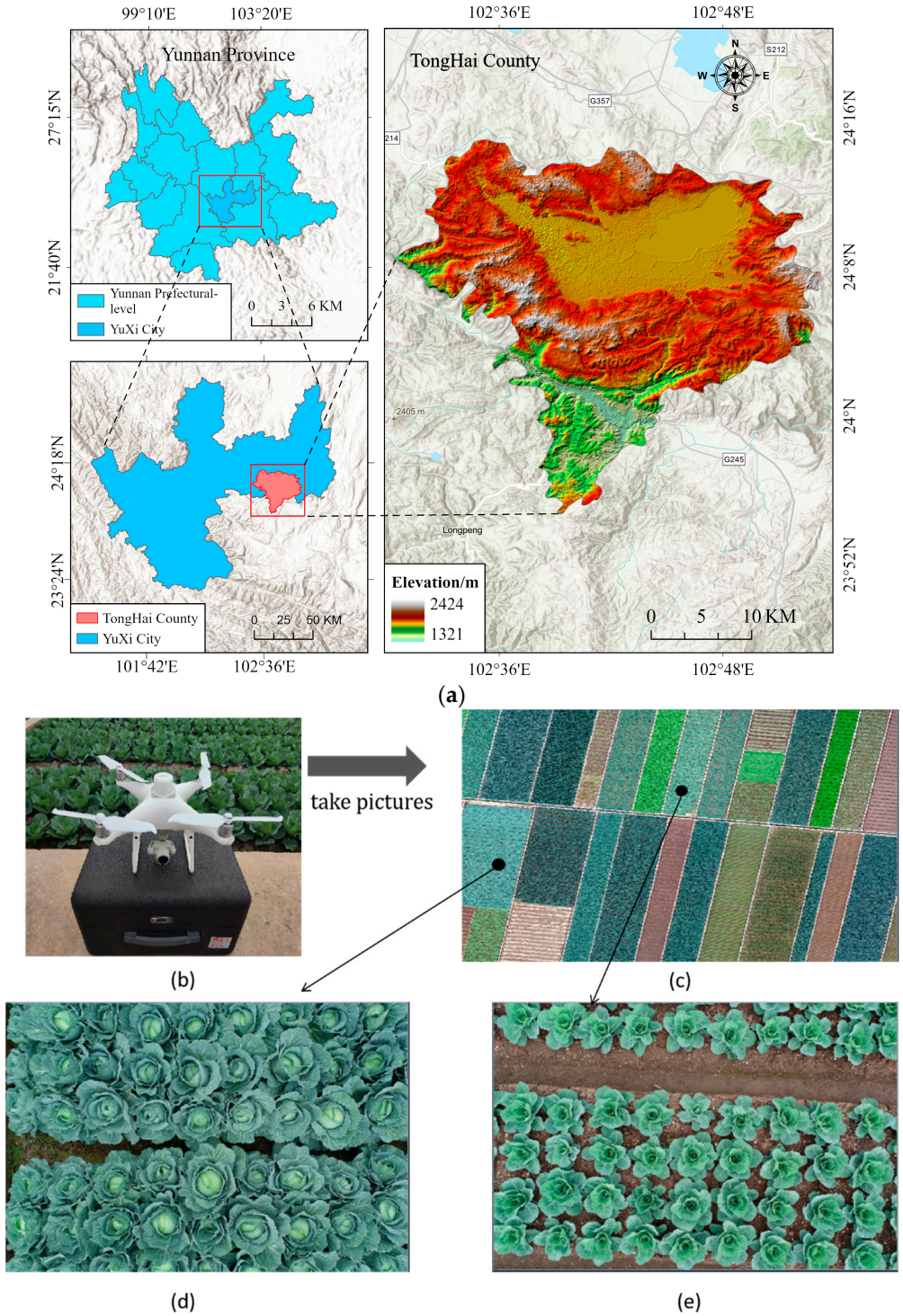


Figure 1. Data acquisition area, cultivation zones, and cabbage sample images. (**a**) Geographic location of the cabbage sampling site; (**b**) UAV platform used for image acquisition; (**c**) Cabbage cultivation area captured at an altitude of 30 m; (**d**,**e**) Cabbage head sample images captured at an altitude of 3 m during different growth stages.

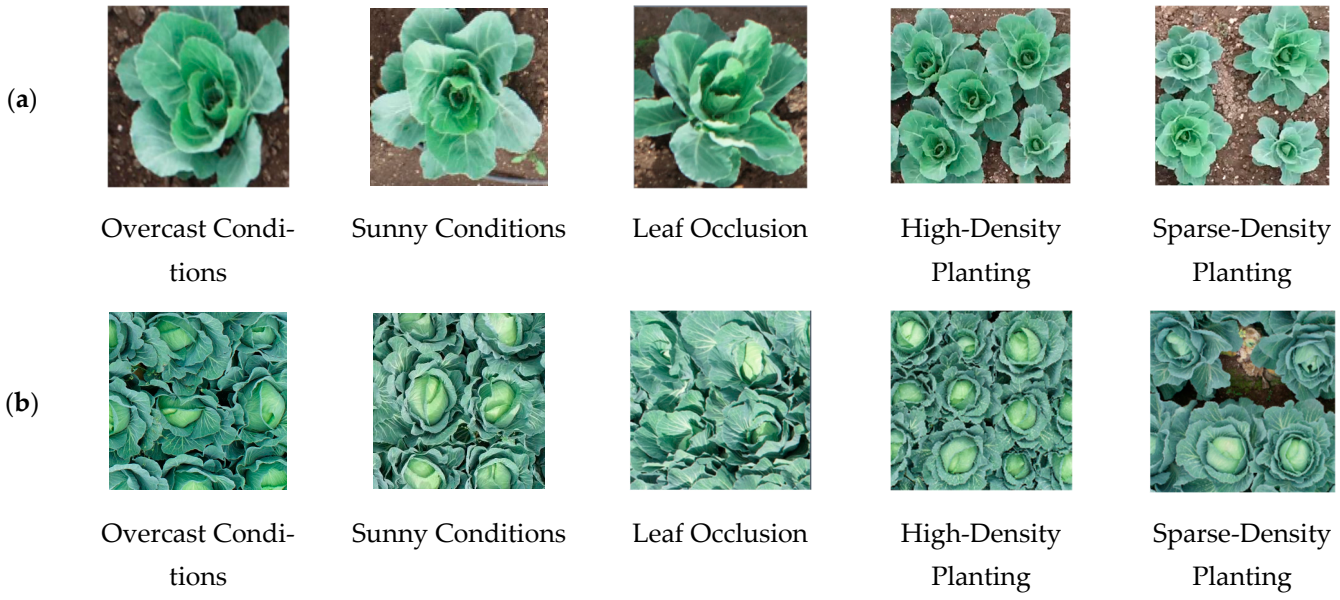


Figure 2. Cabbage sample images under varying complex field conditions. (**a**) Sample image during the early heading stage of cabbage; (**b**) Sample image during the compact heading stage of cabbage.

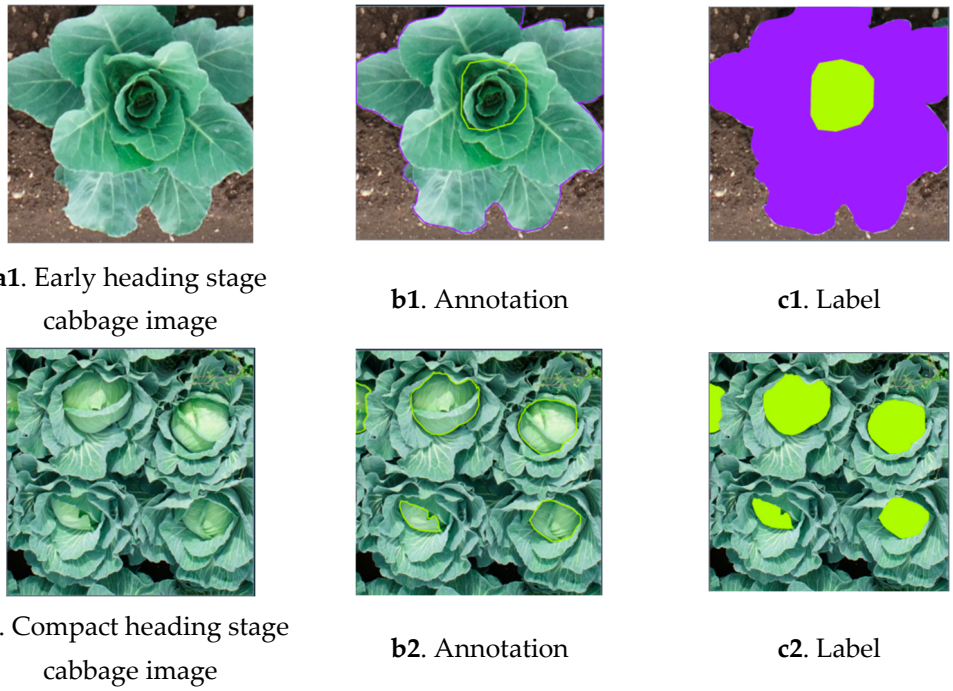


Figure 3. Example of cabbage image annotation. (**a1**–**c1**) represent the annotation process for cabbage images in the early heading stage. (**a2**–**c2**) show the annotation process for images in the compact heading stage.

After augmentation, the number of samples increased to 1602 for the early heading stage and 3354 for the compact heading stage. Subsequently, all image samples were randomly shuffled to ensure independent and identically distributed sampling. The dataset was then split into training, validation, and test sets in a 7:2:1 ratio. This widely adopted split strategy provides a practical balance—offering sufficient data for model training while retaining enough samples in the validation and test sets to ensure statistical robustness and prevent overfitting. Each subset was mutually exclusive, with no overlapping images across partitions. Examples of augmented images are illustrated in Figure 4, and the final sample distribution across the three subsets is summarized in Table 1.

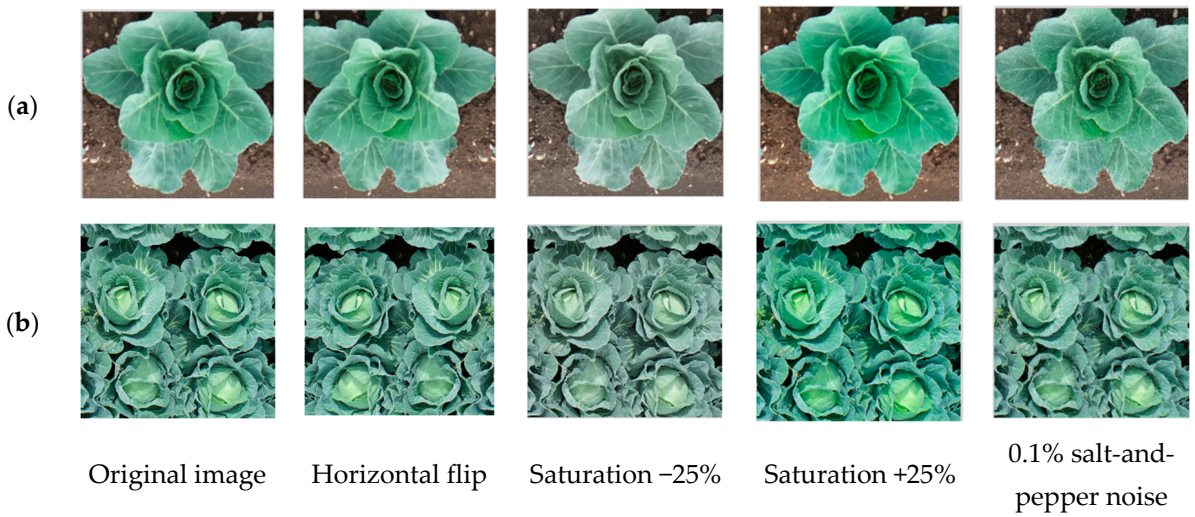


Figure 4. Example of data-augmented cabbage images. (**a**) Sample image during the early heading stage of cabbage; (**b**) Sample image during the compact heading stage of cabbage.

Table 1. Division of the cabbage datasets.

Datasets	Total Number of Pictures	Traning Set	Val Set	Test Set
Early heading stage cabbage image	1602	1101	314	187
Compact heading stage cab-bage image	3354	2347	670	337

2.3. Methods

2.3.1. U-MoEMamba Semantic Segmentation Network

Cabbage image segmentation presents multiple challenges due to variations in lighting, irregular target shapes, dense overlapping leaves, and high inter-class texture similarity. These factors often lead to issues such as segmentation omissions and false positives, demanding enhanced robustness and precision from segmentation models. To tackle these challenges and improve cabbage head segmentation in complex environments, we propose U-MoEMamba. The model integrates the Mixture-of-Experts (MoE) mechanism with the Mamba architecture, aiming to simultaneously capture long-range dependencies and support multi-scale feature fusion, thereby striking a balance between global context understanding and fine-grained local detail extraction.

As illustrated in Figure 5, U-MoEMamba adopts the classic U-Net encoder–decoder architecture. The encoder is based on the lightweight and efficient ResT backbone, which extracts rich multi-scale features. The decoder is deeply customized and optimized with three core components: the MambaMoEFusion module for expert path integration, the MSCrossDualAttention module for enhanced multi-scale context modeling, and the MOE-GatingNetwork, a lightweight adaptive gating unit for expert selection. A key innovation lies in the design of a heterogeneous expert pool composed of three distinct feature processing branches (a multi-scale convolutional path, a cross-attention path, and a Mamba-based state-space path). These experts are dynamically fused through the MOEGatingNetwork, a lightweight gating module that learns to assign optimal weights to each expert output. This design enables adaptive expert selection and context-aware feature fusion, effectively addressing the limitations of conventional single-path feature extractors that often lack flexibility and task-specific adaptability. Through this hybrid design, U-MoEMamba

demonstrates strong capability in handling complex segmentation scenarios, improving both precision and generalization under challenging field conditions.

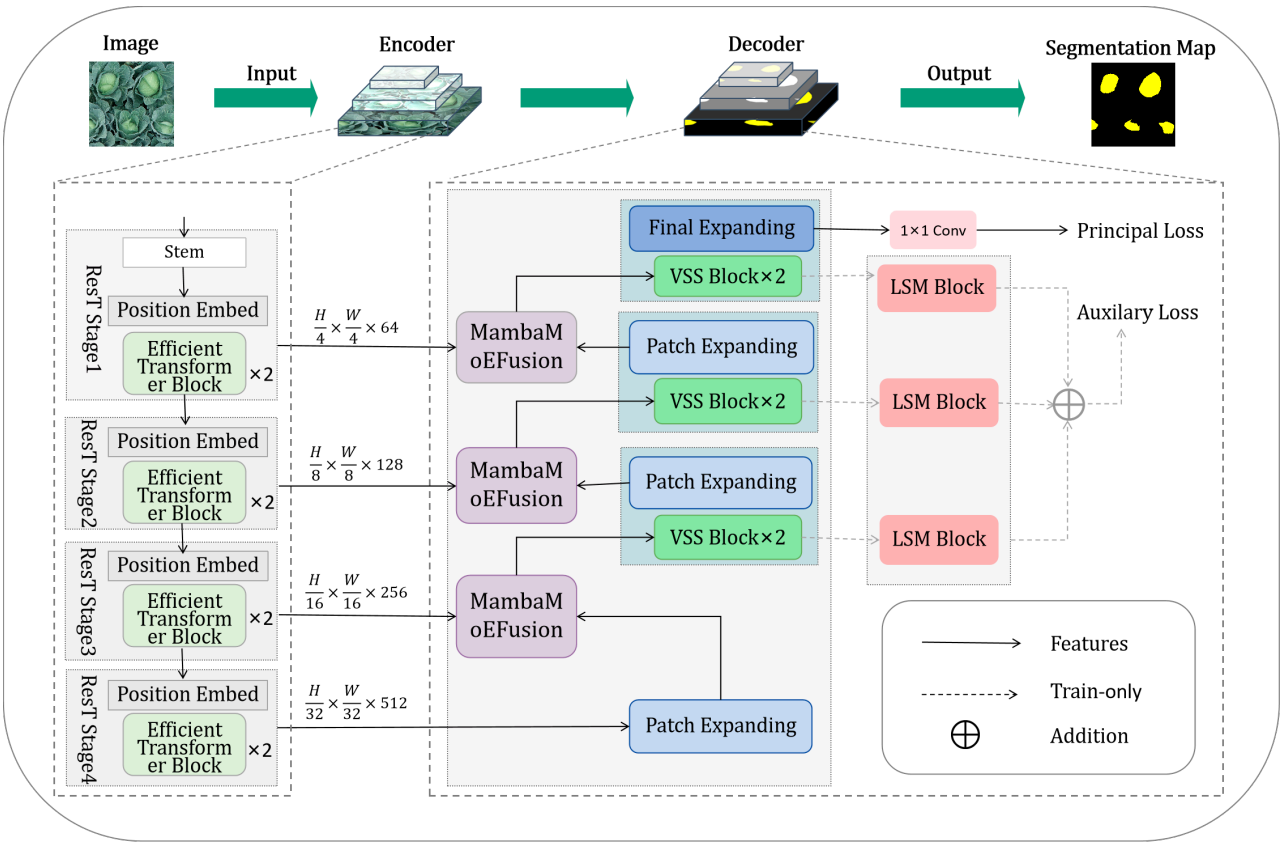


Figure 5. Overall architecture of the U-MoEMamba model.

2.3.2. MambaMoEFusion: Hybrid Expert Gating Fusion Module

Precise segmentation of cabbage in complex field environments is often hindered by challenges such as under-segmentation and mis-segmentation. To address these issues, we propose a novel Hybrid Expert Gating Fusion Module, termed MambaMoEFusion, as a core component of the U-MoEMamba network (see Figure 6 for the architecture). This module integrates three complementary feature extraction paradigms, each acting as an expert within a Mixture-of-Experts (MoE) framework (1) Multi-Scale Convolutional Expert (CNN-based): Efficient at capturing local spatial structures and fine-grained texture details. (2) Attention-Based Expert: Specializes in modeling global contextual dependencies, enhancing the model's awareness of semantic relationships across the image. (3) State-space Model Expert (VSSBlock): Based on Mamba, this expert excels at long-range sequence modeling, enabling the capture of cross-scale spatial dependencies with high computational efficiency.

To effectively integrate these heterogeneous expert pathways, the MambaMoEFusion module employs a learnable gating network, which dynamically computes the contribution weight of each expert based on the input feature context. The outputs of all experts are then adaptively fused through weighted summation, enabling the network to automatically emphasize the most relevant feature representation for each input scenario. This dynamic fusion mechanism harnesses the unique strengths of different paradigms, mitigating the limitations of single-path feature extraction methods.

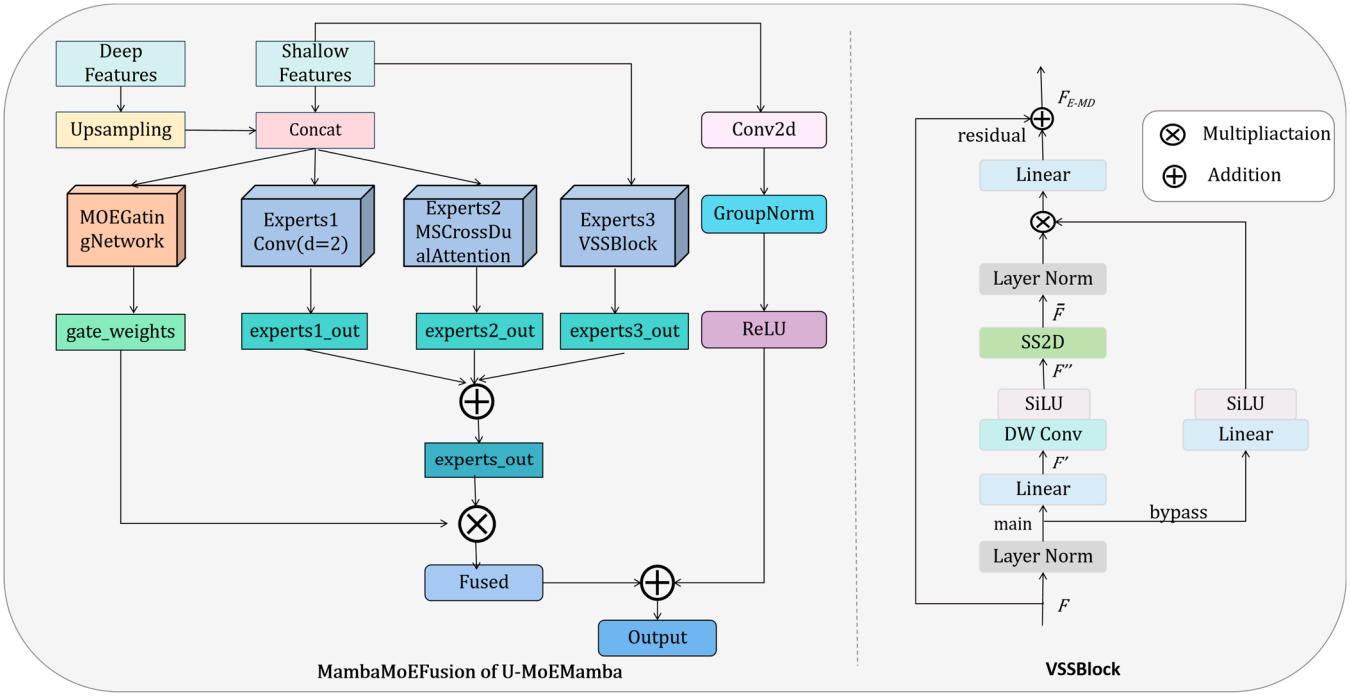


Figure 6. Structure of the MambaMoEFusion module.

The MambaMoEFusion module is applied at each upsampling stage in the decoder, immediately after the PatchExpand operation. At each stage, a deep feature map ($D' \in R^{B\times C_d\times H\times W}$) is first upsampled via bilinear interpolation to match the spatial resolution of the corresponding shallow feature map ($S \in R^{B\times C_s\times H\times W}$). The upsampled deep features (D') are then concatenated with the shallow features along the channel dimension to form a fused representation X_{cat}. For detailed formulations, refer to Equations (1) and (2).

$$D' = Interp(D, size = (H, W)) \tag{1}$$

$$X_{cat} = Concat(S, D') \in R^{B\times(C_s+C_d)\times H\times W} \tag{2}$$

In this context, $S \in R^{B\times C_s\times H\times W}$ denotes the shallow feature map from the encoder, $D' \in R^{B\times C_d\times H\times W}$ represents the deep feature map from the previous decoder layer, D' is the upsampled version of $D \in R^{B\times C_d\times H\times W}$, and X_{cat} is the channel-wise concatenated feature map.

The MambaMoEFusion module incorporates three heterogeneous expert pathways to process the concatenated feature map X_{cat} to produce diverse semantic representations.

Expert 1 is a multiscale convolutional expert consisting of two layers of convolution. The first layer applies a standard 3×3 convolution to the fused input E_1, followed by Group Normalization (GN) with four groups, producing an intermediate output E_1'. The second layer utilizes a dilated 3×3 convolution with a dilation rate of 2 to expand the receptive field and capture wider contextual information without sacrificing resolution. For detailed formulations, refer to Equations (3)–(5).

$$E_1 = Conv_{3\times3}(X_{cat}) \tag{3}$$

$$E_1' = GroupNorm(E_1) \tag{4}$$

$$E_1^{out} = Conv_{3\times3}^{dil=2}(\mathrm{ReLU}(E_1')) \tag{5}$$

where $dil = 2$ denotes a dilation rate of 2.

Expert 2 is a cross-attention expert, which employs a dual-path attention mechanism composed of cross-channel attention and multi-scale spatial attention. It operates on both the shallow feature map (S) and the upsampled deep feature map (D') to enhance feature interaction across semantic levels. First, both S and D' are activated via a sigmoid function, yielding corresponding spatial attention weights A_s and A_d that highlight key semantic regions in the shallow and deep features, respectively. Next, a dilated convolution with dilation rate r is applied to generate a multi-scale spatial attention mask $M_{spatial}$, which captures broader contextual dependencies across the concatenated feature space. Finally, the outputs of the two attention branches are fused through weighted aggregation, producing the final attention-enhanced representation E_2^{out}.For detailed formulations, refer to Equations (6)–(8).

$$A_s = \sigma(Conv_{1\times1}(S)), A_d = \sigma(Conv_{1\times1}(D')) \tag{6}$$

where σ denotes Sigmoid activation and $A^s \in [0,1]^{B\times1\times H\times W}$.

$$M_{\mathrm{sp}atial} = \sigma(\sum_{r\in\{2,4,6\}} DilConv_r(Concat(A_S, A_D))) \tag{7}$$

$$E_2^{out} = M_{spatial} \odot (A_s \otimes S) + (1 - M_{spatial}) \odot (A_d \otimes D') \tag{8}$$

where $DilConv_r$ denotes a dilated convolution operation with dilation rate r and an output channel count of 1. Here, $\odot$ represents element-wise multiplication, while $\otimes$ signifies channel-wise broadcast multiplication.

Expert 3 is a Mamba-path expert that incorporates a state-space model. The process involves the following steps: First, the current input S' is dimensionally rearranged into a sequential format. This input state is then discretized to yield $\overline{A}$. Next, state-space equations are employed for modeling, producing the output y_t . Finally, this output is restored to the original image format. For detailed formulations, refer to Equations (9)–(13).

$$S' = Permute(S) \in R^{B\times H\times W\times C_s} \tag{9}$$

$$\overline{A} = e^{\Delta t A} \tag{10}$$

$$h_t = \overline{A} h_{t-1} + \overline{B}\, s'_t, h_t \in R^D \tag{11}$$

$$y_t = Ch_t \tag{12}$$

$$E_3^{out} = Permute^{-1}(y) \tag{13}$$

where h_t is the hidden state, and D denotes the state dimension. A, B, and C represent learnable parameter matrices. Here, s'_t denotes the input vector at position t.

This heterogeneous expert combination enables comprehensive feature representation spanning from local to global contexts, effectively handling distinct feature paradigms.

Within the MambaMoEFusion module, we designed a lightweight gating network, the MOEGatingNetwork, featuring a bottleneck structure to achieve dynamic expert selection. Based on the input features, the network generates a distribution of expert weights. Feature compression is performed via average pooling, followed by sparse activation realized through bottleneck mapping and probability normalization. For detailed formulations, see Equations (14)–(16).

$$v = GAP(X_{cat}) \in R^{B\times(C_s + C_d)} \tag{14}$$

$$z = W_2 \cdot ReLU(W_1 \cdot v + b_1) + b_2 \tag{15}$$

$$g = Softmax(z) \in R^{B\times N_e} \tag{16}$$

Here, *GAP* denotes the global average pooling operation applied to the fused features, producing the intermediate vector $v \in R^{B\times(C_S+C_d)}$. Z represents the gating scores learned by the gating network. $W_1 \in R^{(C_s+C_d)\times C_m}$ and $W_2 \in R^{N_e\times C_m}$ denote the weight matrices of the first and second fully connected layers, respectively, $C_m = \max(4, (C_s + C_d) \div 4)$ indicates the channel compression ratio, and $b_1 \in R^{C_m}$ is the bias term of the first layer. $b_2 \in R^{N_e}$ corresponds to the bias term of the second layer.

Each expert output is weighted by its corresponding gating score, and the weighted sum is passed through a convolutional layer to enhance feature representation, producing an intermediate tensor $Y_{enhance}$. Next, a residual connection is employed by adding $Y_{enhance}$ to the original shallow feature map, thereby preserving the initial information flow and yielding the final output Y_{out}, as defined in Equations (17)–(19):

$$Y_{fused} = \sum_{i=1}^{Ne} g_i \cdot E_i^{out}, g_i \in g, Y_{fused} \in R^{B\times C_s \times H\times W} \tag{17}$$

Here, $g = [g_1, g_2, g_3]$ denotes the gating weight vector.

$$Y_{enhance} = Conv_{3\times3}(Y_{fused}) \tag{18}$$

$$Y_{out} = GroupNorm(ReLU(Y_{enhance})) + S \tag{19}$$

This mechanism enables the expert to effectively emphasize relevant spatial structures and contextual cues, improving segmentation performance in regions affected by occlusion, background clutter, or weak object boundaries.

2.3.3. MSCrossDualAttention: Multi-Scale Cross Dual Attention Module

The cross-attention mechanism effectively facilitates the interaction between shallow features, which encode spatial details, and deep features, which capture high-level semantic information. Meanwhile, multi-scale convolution excels at extracting features from objects of varying sizes, such as cabbages at different growth stages. To synergistically exploit the complementary strengths of these two paradigms, we propose an innovative module named Multi-Scale Cross Dual Attention (MSCrossDualAttention), which is integrated into the network as a dedicated expert branch. By combining cross-attention with multi-scale feature processing, this module significantly enhances the network's ability to model multi-scale targets and complex semantic dependencies. The detailed architecture of the MSCrossDualAttention module is illustrated in Figure 7.

Within the cross-channel attention path, the deep feature map (D) is first upsampled via bilinear interpolation to match the spatial resolution of the shallow features. Based on this alignment, two attention weight maps are generated: A_s for the shallow feature map and A_d for the upsampled deep feature map. Next, a feature decoupling operation is performed by applying element-wise multiplication between the original features and their respective attention weights. This selectively enhances spatially significant regions, as defined in Equations (20)–(22).

$$A_s = \sigma(Conv_{1\times1}(S)), A_d = \sigma(Conv_{1\times1}(D')) \tag{20}$$

$$Senh = A_s \odot S \tag{21}$$

$$D_{enh}{}' = A_d \odot D' \tag{22}$$

In the multi-scale spatial attention path, the attention maps A_s and A_d are concatenated along the channel dimension to form an integrated feature map M_{cat}. A set of dilated convolutions with varying dilation rates is then applied to capture multi-scale contextual

dependencies. The outputs of these convolutions are concatenated and passed through a 1×1 convolution to compress the representation into a single-channel spatial attention mask. For detailed formulations, refer to Equations (23) and (24).

$$M_{cat} = Concat(A_s, A_d) \in R^{B \times 2 \times H \times W} \tag{23}$$

$$M_{spatial} = \sigma(Conv_{1 \times 1}(Concat(M_{rate=2}, M_{rate=4}, M_{rate=6}))) \tag{24}$$

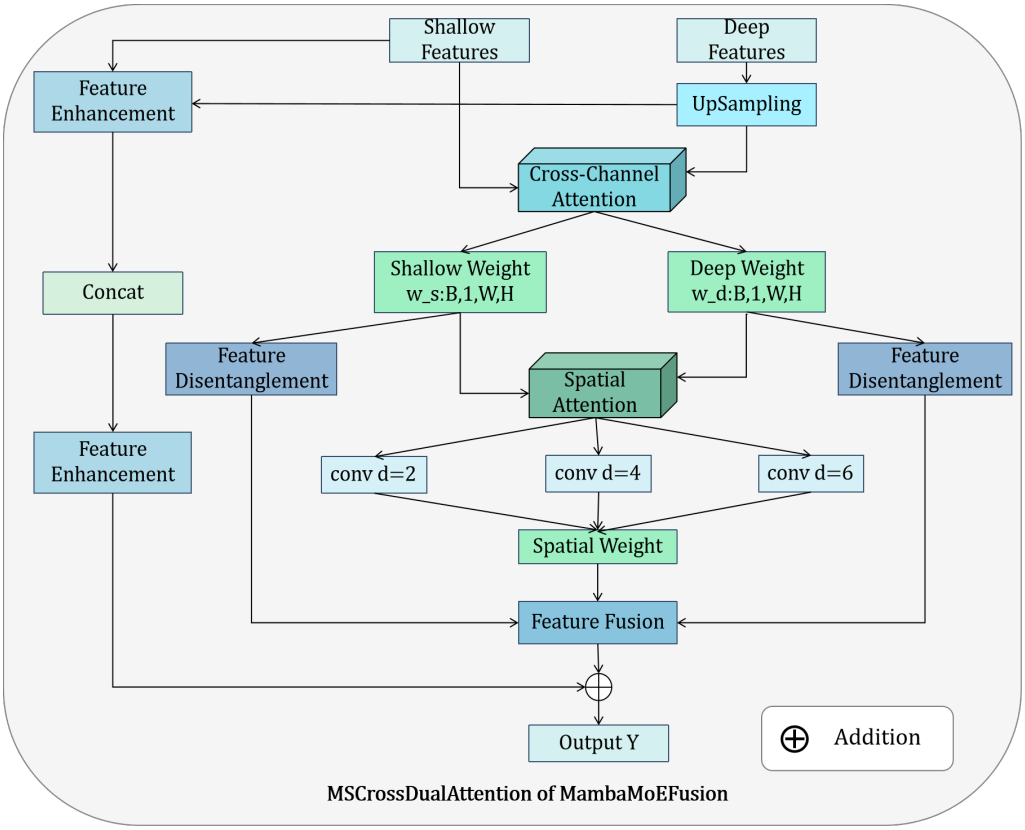


Figure 7. Structure of the MSCrossDualAttention module.

Finally, the outputs from both the cross-channel attention and the multi-scale spatial attention paths are fused using the learned spatial mask, enabling dynamic, location-aware weighting of features. The original shallow features and the upsampled deep features are also concatenated and enhanced through a group of convolutional layers, followed by a residual connection that adds the enhanced features to the fused output, yielding the final refined representation. For detailed formulations, refer to Equations (25)–(27).

$$F_{fused} = M_{spatial} \odot S_{enh} + (1 - M_{spatial}) \odot D_{enh}' \tag{25}$$

$$E_{enh} = ReLU(GroupNorm(Conv_{3 \times 3}(Concat(S, D')))) \tag{26}$$

$$Y_{out} = F_{fused} + E_{enh} \tag{27}$$

2.3.4. MOEGatingNetwork: Gating Fusion Module

Given the limitations of the baseline gating mechanism in the original MambaMoEFusion module regarding feature adaptability, this study enhances the module by introducing a lightweight dynamic-routing gating fusion module (MOEGatingNetwork) to improve cabbage segmentation accuracy in complex field scenarios. The key innovation lies in its ability to adaptively generate optimal expert weight distributions based on the intrinsic

characteristics of the input features. The architecture of the MOEGatingNetwork module is illustrated in Figure 8.

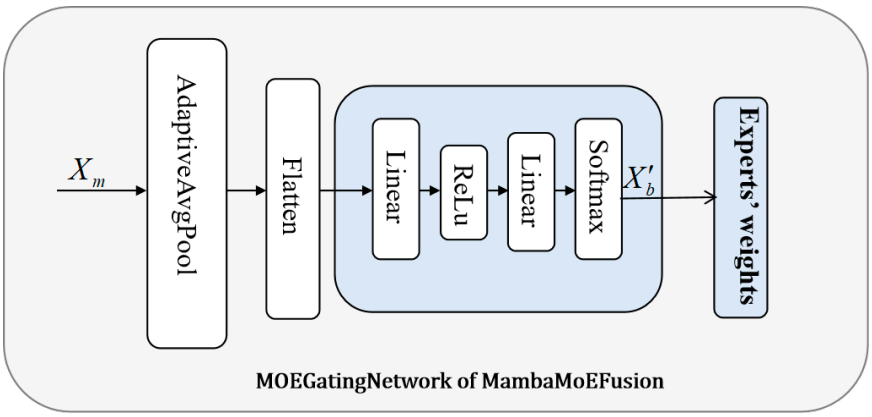


Figure 8. Structure of the MOEGatingNetwork module.

Specifically, the input feature map $\left(x \in R^{B\times C\times H\times W}\right)$ is first passed through a global average pooling layer to produce a compact vector representation ($v \in R^{B\times C}$), which suppresses spatial redundancy while preserving channel-wise statistics. Next, a fully connected layer reduces the channel dimension from C to a compressed size $C_m = max(4, c/4)$, effectively reducing the parameter complexity. A ReLU activation function is applied to introduce non-linearity and enhance feature expressiveness. The resulting vector is then projected into a score vector representing the matching strength for each expert. Finally, a softmax normalization is applied to convert the scores into a probability distribution over the expert paths, yielding the final expert weight vector.

This dynamic and data-dependent routing strategy enables the network to selectively emphasize the most suitable expert pathways, thereby improving its adaptability and robustness in complex segmentation tasks. For detailed formulations, refer to Equations (28)–(30).

$$z = ReLU(W_1 v + b_1), W_1 \in R^{c\times c_m} \tag{28}$$

$$s = W_2\, z + b_2\,, W_2\ \in R^{C_m\ \times N_e} \tag{29}$$

$$g = softmax(s) \in R^{B\times N_e} \tag{30}$$

To further optimize training, a load-balancing strategy was employed to ensure equitable expert utilization. Within the MOEGatingNetwork, a sparse activation mechanism was implemented, whereby only the expert with the highest weight is activated. This significantly reduces computational overhead while maintaining segmentation performance, as defined in Equation (31).

$$Selected\ Experts = arg \underset{i\in[1,N_e]}{\overset{(K)}{\max}} g_i \tag{31}$$

2.4. Experimental Platform and Parameter Settings

All experiments in this study were implemented using the PyTorch 2.5.1 deep learning framework. The hardware platform utilized an NVIDIA RTX A4000 GPU (16 GB VRAM), running on an Ubuntu 20.04.6 operating system within a Python 3.8 software environment. To ensure reproducibility and fairness, all experiments were conducted under identical training and testing configurations: a batch size of 8, the Adam optimizer with an initial learning rate of 6×10^{-4}, and a total training duration of 100 epochs. A three-tiered loss system was employed to comprehensively optimize model training: the Primary

Loss utilized a combined Cross-Entropy and Dice loss for core segmentation optimization; the Auxiliary Loss incorporated supervision signals at intermediate layers to enhance hierarchical feature learning; and the Expert Load Balancing Loss mitigated the "Winner-Takes-All" phenomenon among experts, thereby improving model capacity utilization and ensuring even computational load distribution.

2.5. Evaluation Metrics

To comprehensively evaluate the segmentation performance of the proposed model on the cabbage dataset, and given that the task is a pixel-level classification problem involving three semantic classes (cabbage head, leaves, and background), we selected five key metrics that are widely adopted in semantic segmentation research: GFLOPS (computational efficiency), Params (parameter count), F1 score (the harmonic mean of precision and recall), mIoU (mean Intersection over Union), and OA (Overall Accuracy). These metrics collectively capture both regional consistency and pixel-wise classification performance. The evaluation was conducted systematically from three perspectives—computational efficiency, model complexity, and segmentation accuracy—to ensure the scientific rigor and validity of the experimental results.

Specifically, GFLOPS and Params quantify the model's scale and computational complexity; the F1 score, defined as the harmonic mean of precision and recall, effectively measures the model's ability to balance false positives and false negatives in classification tasks; IoU assesses the segmentation precision for each class by computing the ratio of the intersection to the union between predicted and ground truth masks, and mIoU averages these across all classes to reflect overall segmentation consistency; and OA quantifies the proportion of correctly predicted pixels relative to the total pixel count, directly reflecting the model's overall segmentation capability. For detailed formulations, refer to Equations (32)–(37).

$$precision = \frac{TP}{TP + FP} \tag{32}$$

$$Recall = \frac{TP}{TP + FN} \tag{33}$$

$$F1Score = \frac{precision \times Recall}{precision + Recall} \tag{34}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{35}$$

$$mIoU = \frac{1}{N}\sum_{i=1}^{n} IoU_i \tag{36}$$

$$OA = \frac{TP + FN}{TP + FN + FP + TN} \tag{37}$$

Here, TP denotes correctly identified target pixels, FP represents background pixels misclassified as targets, FN indicates target pixels erroneously classified as background, and TN corresponds to correctly identified background pixels.

3. Results

To comprehensively evaluate the model performance, we conducted systematic comparative experiments using several state-of-the-art Mamba-based segmentation models on two distinct datasets: one representing the early heading stage and the other the compact heading stage of cabbage growth. Experimental results demonstrate that the proposed U-MoEMamba model consistently achieves superior performance across multiple key evaluation metrics on both datasets, outperforming all baseline methods in the comparison.

Further ablation studies validate the effectiveness of the heterogeneous expert collaboration architecture embedded within U-MoEMamba, highlighting its contribution to improved segmentation accuracy and model robustness in complex agricultural environments.

3.1. Comparative Analysis on the Early Heading Stage of Cabbage

To assess the robustness of the proposed model under varying environmental conditions, experiments were first conducted on the early heading stage cabbage dataset, which encompasses a wide range of typical field scenarios, including varying lighting conditions, planting densities, and degrees of leaf occlusion. Quantitative analysis reveals that the proposed method consistently outperforms existing approaches across multiple key evaluation metrics. As shown in Table 2, we systematically compared U-MoEMamba with several mainstream Mamba-based segmentation models, including SegMamba (pure-Mamba architecture), MambaUNet (UNet-integrated architecture), CM-UNet (convolution-enhanced architecture), RS3-Mamba (residual-optimized architecture), and EfficientPyramidMamba (efficient-pyramid design).

Table 2. Validation results of different segmentation models on the early heading stage cabbage dataset. Bold values indicate the best performance for each metric.

Method	Backgrounds		Cabbage Fruit		Cabbage Leaves		mIoU (%)	mF1 (%)	OA (%)
	IoU (%)	F1 (%)	IoU (%)	F1 (%)	IoU (%)	F1 (%)			
SegMamba	95.21	97.54	65.47	79.13	96.11	98.02	85.60	91.56	97.30
MambaUnet	95.10	97.49	72.61	84.13	96.47	98.20	88.06	93.26	97.55
CM-UNet	95.45	97.67	71.76	83.56	96.58	98.26	87.93	93.16	97.63
RS3-Mamba	95.44	97.67	72.20	83.85	96.55	98.24	88.06	93.25	97.61
EfficientPyramidMamba	95.08	97.48	72.76	84.23	96.49	98.21	88.11	93.30	97.57
U-MoEMamba	**95.68**	**97.79**	**75.96**	**86.63**	**96.90**	**98.42**	**89.51**	**94.28**	**97.85**

In the task of segmenting three target classes—background, cabbage head, and cabbage leaf—U-MoEMamba achieved the highest scores in both IoU and F1 metrics for all categories. Notably, the model demonstrated substantial performance gains in segmenting the two key target structures. The cabbage leaf achieved an IoU of 96.90%, representing a 0.32% improvement over the next best model, with a corresponding F1 score of 98.42%; the cabbage head achieved an IoU of 75.96%, exceeding the second-best result by 3.20%, with an F1 score of 86.63%. In the complex background class, the IoU reached 95.68%, a 0.23 percentage point improvement over CM-UNet (95.45%), with a corresponding F1 score of 97.79%. In terms of overall performance, U-MoEMamba demonstrated clear advantages over existing methods, achieving a mean Intersection-over-Union (mIoU) of 89.51%, which is 1.40% higher than the runner-up EfficientPyramidMamba. The mean F1 score (mF1) reached 94.28%, improving by 1.02%, and the overall accuracy (OA) reached 97.85%, a 0.22% gain over the second-best result. These results collectively demonstrate the superior segmentation accuracy and robustness of U-MoEMamba across all core evaluation metrics under challenging and diverse field conditions.

The U-MoEMamba model achieves a significant breakthrough in the joint optimization of efficiency and performance. As shown in Table 3, with only 20.88 MB of parameters and 165.72 GFLOPs of computational cost, the model achieves a remarkable mIoU of 89.51% on the early heading stage cabbage dataset, setting a new benchmark in segmentation performance. Specifically, in terms of parameter efficiency, U-MoEMamba reduces the model size by 51.8% compared to RS3-Mamba (from 43.32 MB to 20.88 MB), while simultaneously improving mIoU by 1.45 percentage points (from 88.06% to 89.51%). Regarding the trade-off between computational complexity and accuracy, although U-MoEMamba incurs higher FLOPs than the lightweight CM-UNet (165.72 GFLOPs vs. 51.76 GFLOPs), it delivers a substantial mIoU gain of 1.58% (from 87.93% to 89.51%). In terms of overall

segmentation performance, U-MoEMamba consistently outperforms all baseline models, surpassing SegMamba by 3.91% (85.60% → 89.51%), MambaUNet by 1.45%, and EfficientPyramidMamba by 1.40%. These results demonstrate that through its collaborative heterogeneous expert architecture, U-MoEMamba achieves state-of-the-art segmentation accuracy under controllable model complexity, offering an efficient and scalable solution for robust crop identification in complex field environments.

Table 3. Comparison of model parameters, computational cost, and training efficiency on the early heading stage cabbage dataset. Bold values indicate the best performance for each metric.

Method	Params/MB	FLOPs/GB	mIoU
SegMamba	**5.05**	72.99	85.60
MambaUnet	13.88	51.86	88.06
CM-UNet	13.41	**51.76**	87.93
RS3-Mamba	43.32	158.24	88.06
EfficientPyramidMamba	28.76	76.22	88.11
U-MoEMamba	20.88	165.72	**89.51**

Figure 9 presents a qualitative comparison of segmentation results between the proposed U-MoEMamba model and several representative Mamba-based baselines on the early heading stage cabbage dataset. The evaluation covers five representative field scenarios, including overcast conditions (low illumination), sunny conditions (high exposure), leaf occlusion (weed interference), high-density planting (target adhesion), and sparse planting. As illustrated in Figure 9a,b, under extreme lighting conditions such as overcast and bright sunlight, U-MoEMamba produces smoother and more continuous segmentation boundaries that closely align with the ground-truth labels, substantially reducing boundary jaggedness and false segmentation noise. Figure 9c demonstrates the model's capacity to accurately distinguish between occluding weeds and cabbage head structures in cluttered scenes, preserving the morphological integrity of the targets (e.g., complete leaf ball contours). The segmentation results exhibit near-perfect visual consistency with manual annotations. Figure 9d,e further show that in both high- and low-density planting scenarios, baseline models such as SegMamba, CNN-Mamba UNet, and RS3-Mamba commonly suffer from issues such as boundary blurring (e.g., merged adjacent targets) and inter-class confusion (e.g., misclassifying cabbage heads as background). In contrast, U-MoEMamba, powered by its heterogeneous expert collaboration mechanism, maintains the topological continuity of object boundaries and enhances segmentation robustness in complex spatial distributions.

(**a**) Overcast Conditions

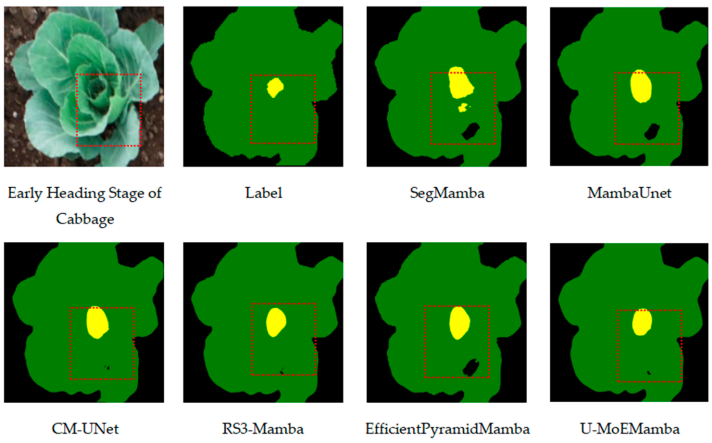


Figure 9. *Cont.*

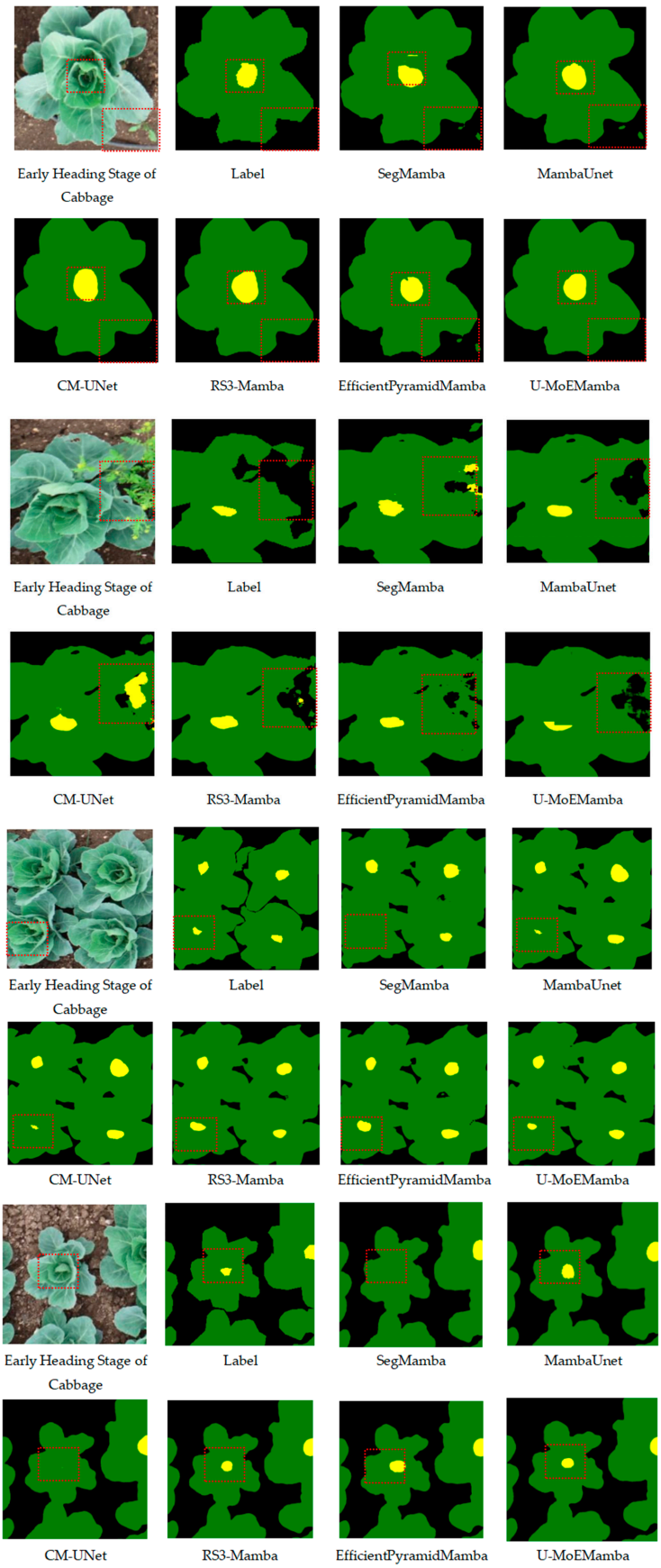


Figure 9. Comparison of segmentation performance during the early heading stage of cabbage using various mainstream Mamba-based networks. The red dotted boxes highlight the visual differences among the results generated by different models.

3.2. Comparative Analysis on the Compact Heading Stage of Cabbage

Experiments conducted on the compact heading stage cabbage dataset further validate the superior performance of the proposed U-MoEMamba model. As shown in Table 4, the model demonstrates clear advantages across both overall performance metrics and class-specific segmentation accuracy. It achieves a mean Intersection-over-Union (mIoU) of 91.88%, surpassing the next best model, CM-UNet (90.98%), by 0.90 percentage points. Similarly, it attains a mean F1 score (mF1) of 95.67%, improving by 0.42 percentage points, and an overall accuracy (OA) of 97.98%, with a 0.22 percentage point gain. For the challenging class of cabbage heads, where inter-class confusion is common, U-MoEMamba achieves an IoU of 86.07%, outperforming RS3-Mamba (85.28%) by 0.79 percentage points, and an F1 score of 92.51%, a 0.46 percentage point improvement over RS3-Mamba (92.05%). The model also maintains stable performance on background segmentation, achieving an IoU of 97.70%, which is 0.73 percentage points higher than that of SegMamba (96.97%).

Table 4. Validation results of different segmentation models on the compact heading stage cabbage dataset. Bold values indicate the best performance for each metric.

Method	Background		Cabbage Fruit		mIoU (%)	mF1 (%)	OA (%)
	IoU (%)	F1 (%)	IoU (%)	F1 (%)			
SegMamba	96.97	98.46	81.98	90.09	89.47	94.28	97.34
MambaUnet	97.35	98.65	84.05	91.33	90.70	96.44	97.67
CM- UNet	97.44	98.70	84.52	91.61	90.98	95.15	97.76
RS3-Mamba	97.56	98.76	85.28	92.05	90.52	95.40	97.86
EfficientPyramidMamba	97.24	98.60	83.22	90.84	90.23	94.72	**98.60**
U-MoEMamba	**97.70**	**98.83**	**86.07**	**92.51**	**91.88**	**95.67**	97.98

These results highlight U-MoEMamba's robust segmentation capabilities across various categories. Particularly noteworthy is its strong performance in segmenting the visually similar cabbage head structures, achieving an IoU of 86.07%, which underscores the effectiveness of the heterogeneous expert architecture in addressing prevalent challenges in agricultural scenes, such as scale variation and inter-class ambiguity.

As shown in Table 5, the U-MoEMamba model also achieves a remarkable balance between accuracy and complexity on the compact heading stage cabbage dataset. With a lightweight architecture consisting of only 20.88 MB of parameters, the model significantly reduces the parameter size by 51.8% compared to the heavier RS3-Mamba (43.32 MB), while attaining a state-of-the-art mIoU of 91.88%, which surpasses the next-best model, CM-UNet, by 0.90 percentage points. These results further validate the effectiveness of the heterogeneous expert architecture in agricultural visual recognition tasks.

Table 5. Comparison of model parameters, computational cost, and training efficiency on the compact heading stage cabbage dataset. Bold values indicate the best performance for each metric.

Method	Params/MB	FLOPs/GB	mIoU
SegMamba	**5.05**	72.99	89.47
MambaUnet	13.88	51.86	90.70
CM- UNet	13.41	**51.76**	90.98
RS3-Mamba	43.32	158.24	90.52
EfficientPyramidMamba	28.76	76.22	90.23
U-MoEMamba	20.88	165.72	**91.88**

Figure 10 presents the segmentation visualization results of U-MoEMamba across five challenging field scenarios during the compact heading stage of cabbage growth, including cloudy conditions, intense sunlight, leaf occlusion, high-density planting, and sparse plant-

ing. These results further validate the model's robustness across diverse environmental conditions. As shown in Figure 10a,b, under extreme lighting variations, U-MoEMamba produces more structurally complete segmentation outputs than competing models, with higher semantic boundary resolution and a marked reduction in edge discontinuities caused by over- or under-exposure. Figure 10c illustrates that in occluded environments, baseline models such as CM-UNet and EfficientPyramidMamba tend to suffer from semantic misclassification (e.g., mislabeling cabbage leaves as cabbage heads) and boundary ambiguity (e.g., diffusion of head contours). In contrast, U-MoEMamba effectively isolates the occluded regions and generates segmentation masks that closely align with manual annotations, preserving the morphological integrity of the target objects. Figure 10d,e further demonstrate that in both densely and sparsely planted scenes, U-MoEMamba produces smoother and more precise boundary contours compared to other models.

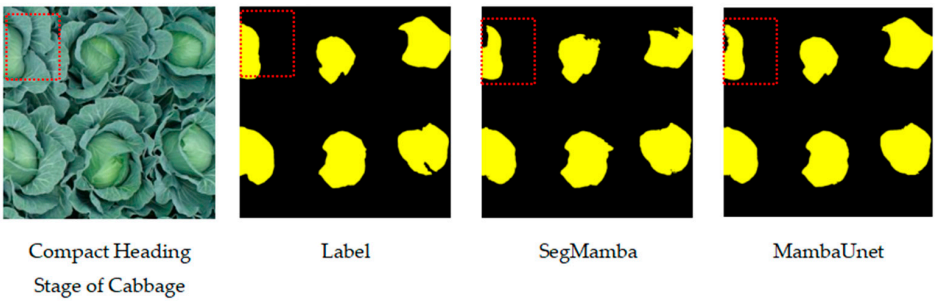
These findings reinforce the conclusions drawn from the early heading stage dataset, offering cross-stage validation of the model's generalization capability. The results highlight that the MoE-based dynamic expert selection mechanism can adaptively activate the most appropriate feature extractor based on scene complexity (e.g., favoring the CNN expert under occlusion), while the long-range modeling capacity of the Mamba architecture ensures the preservation of topological continuity under varying planting densities. Together, these components form the foundation of U-MoEMamba's superior cross-scenario generalizability.

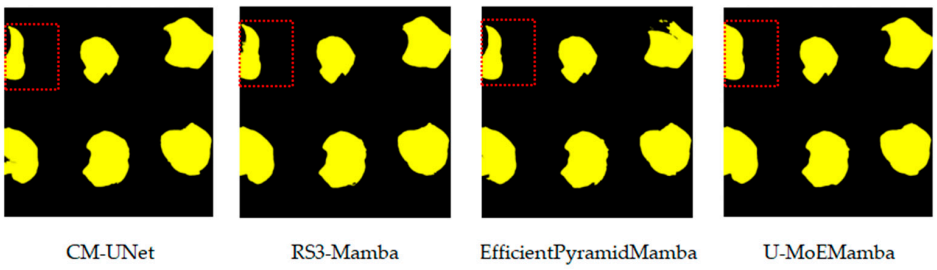
In low-altitude UAV remote sensing scenarios under complex field conditions, U-MoEMamba achieved a 1.4–3.91% improvement in mIoU on the early heading stage cabbage dataset and a 0.9–2.41% gain on the compact heading stage dataset compared to other mainstream Mamba-based models. This performance enhancement holds significant practical value in real-world agricultural applications. In high-throughput field phenotyping and precision agriculture tasks, improved segmentation accuracy can substantially reduce the omission and misclassification of cabbage heads, enabling more reliable crop counting, biomass estimation, and growth monitoring. These improvements support optimized decision-making in fertilization, irrigation scheduling, and yield prediction. Moreover, higher segmentation precision reduces the need for manual correction, lowering labor costs and enhancing operational efficiency in large-scale agricultural deployments. Therefore, the accuracy advantage of U-MoEMamba not only improves the reliability of segmentation tasks but also provides strong support for the practicality and scalability of intelligent agricultural management systems.

3.3. Ablation Study

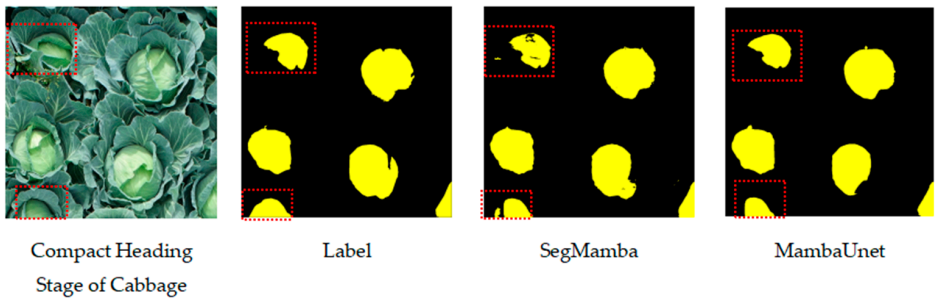
To evaluate the effectiveness of the proposed heterogeneous expert collaboration architecture and its core components, a comprehensive ablation study was conducted on both the early heading and compact heading stage cabbage datasets. Using ResT-Lite as the backbone network, we progressively integrated different expert modules into the decoder—namely, Expert 1 (Multi-scale Convolution), Expert 2 (MSCrossDualAttention), and Expert 3 (VSSBlock for state-space modeling)—as well as the gating fusion module (MOEGatingNetwork). The impact of each module on overall model performance was then systematically assessed. Detailed results are presented in Tables 6 and 7 for the early and compact heading stages, respectively. Evaluation metrics include the number of parameters (Params), computational complexity (FLOPs), class-wise Intersection over Union (IoU), and mean IoU (mIoU), providing a comprehensive understanding of how each component contributes to the model's segmentation performance.

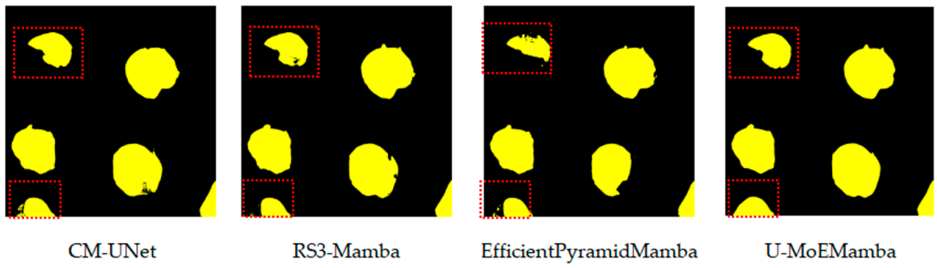
(**a**) Overcast Conditions



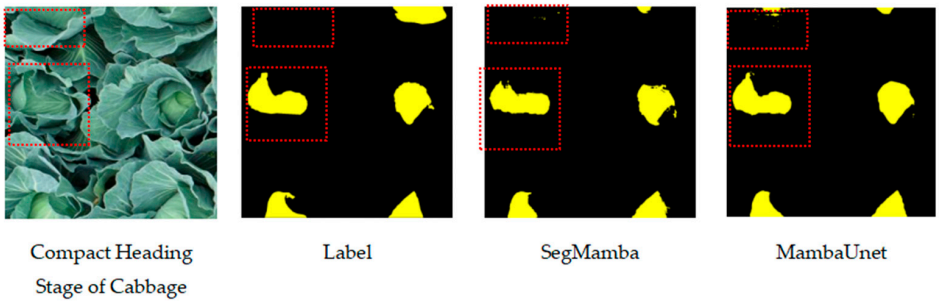


(**b**) Sunny Conditions





(**c**) Leaf Occlusion



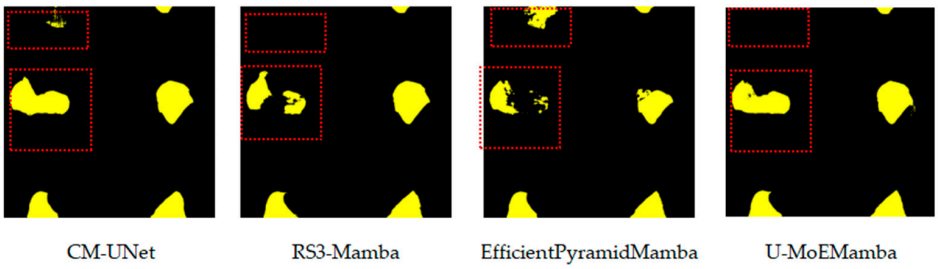


Figure 10. *Cont.*

(**d**) High-Density Planting

(**e**) Sparse-Density Planting

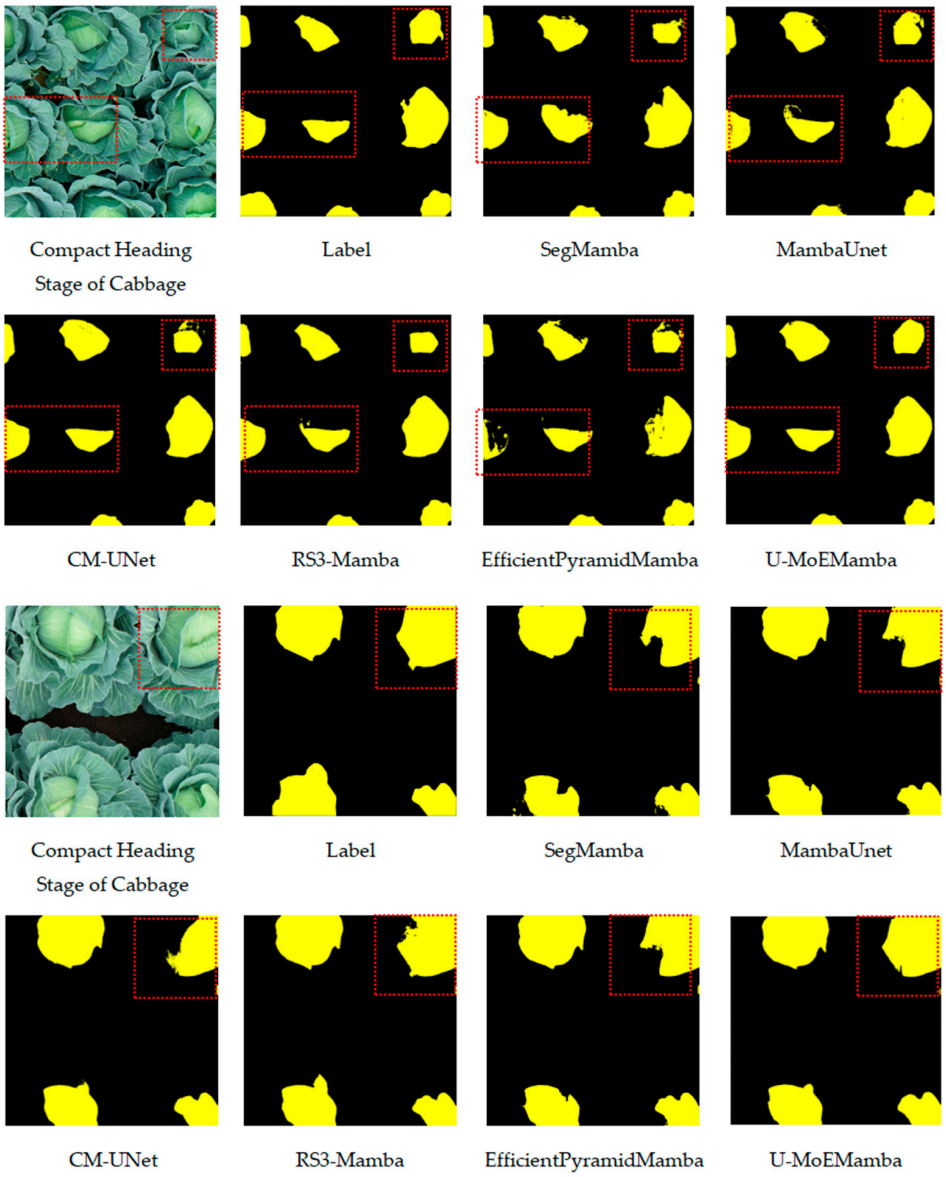


Figure 10. Comparison of segmentation performance during the compact heading stage of cabbage using various mainstream Mamba-based networks. The red dotted boxes highlight the visual differences among the results generated by different models.

Table 6. Ablation study results on the early heading stage cabbage dataset. Bold values indicate the best performance for each metric. "√" denotes the progressive addition of expert modules during the ablation experiments.

Expert 1 (Multi-Scale Conv)	Expert 2 (MSCrossDualAttention)	Expert 3 (VSSBlock)	MOEGatingNetwork	Param (M)	FLOPS (G)	Iou (%)			mIoU (%)
						Backgrounds	Cabbage Fruit	Cabbage Leaves	
				16.90	128.80	95.44	72.20	96.55	88.06
√				17.68	136.09	95.64	74.44	96.78	88.95
	√			16.99	129.62	95.63	74.53	96.79	88.98
		√		**16.84**	**128.48**	95.52	74.84	96.76	89.04
√	√			19.40	151.48	95.65	74.79	96.81	89.08
√		√		19.25	150.34	95.60	75.22	96.50	89.01
	√	√		18.56	143.90	95.65	74.56	96.79	89.00
√	√	√		20.80	165.72	95.62	74.89	96.80	89.10
√	√	√	√	20.88	165.72	**95.68**	**75.96**	**96.90**	**89.51**

Table 7. Ablation study results on the compact heading stage cabbage dataset. Bold values indicate the best performance for each metric. "√" denotes the progressive addition of expert modules during the ablation experiments.

Expert 1 ((Multi-Scale Conv)	Expert 2 (MSCrossDualAttention)	Expert 3 (VSSBlock)	MOEGatingNetwork	Param (M)	FLOPS (G)	Iou (%)		mIoU (%)
						Backgrounds	Cabbage Fruit	
				16.90	128.80	97.59	85.42	91.50
√				17.68	136.09	97.65	85.90	91.77
	√			16.99	129.62	97.19	86.07	91.63
		√		**16.84**	**128.48**	97.62	85.67	91.64
√	√			19.40	151.48	97.55	85.32	91.43
√		√		19.25	150.34	97.51	85.17	91.34
	√	√		18.56	143.90	97.65	85.82	91.73
√	√	√		20.80	165.72	97.63	85.74	91.69
√	√	√	√	20.88	165.72	**97.70**	**86.11**	**91.88**

The individual contributions of each expert module were first evaluated. Results indicate that Expert 1 (Multi-scale Convolution) improved the fruit IoU by 0.23% (from 96.55% to 96.78%) at the early heading stage, with only a 4.6% increase in parameters (to 17.68 M), confirming the advantage of multi-scale convolution for capturing fine fruit details. Expert 2, the novel attention module (MSCrossDualAttention) designed in this study, achieved the largest leaf IoU improvement at the early stage, increasing it by 2.33% to 74.53%, while using fewer parameters (at 16.99 M), highlighting the attention mechanism's strength in modeling long-range dependencies. Expert 3 (VSSBlock) enhanced fruit IoU by 0.25% (85.42% to 85.67%) during the compact heading stage, demonstrating the state-space model's (SSM) advantage in capturing spatial sequential dependencies, with an early-stage leaf IoU of 74.84%, the best among single experts.

Subsequently, integrating any two experts (Expert 1 + Expert 2, Expert 1 + Expert 3, Expert 2 + Expert 3) yielded improved mean IoU values (89.08%, 89.01%, and 89.00%, respectively) at the early heading stage, indicating strong complementarity between expert modules. On the compact heading dataset, dual-expert combinations showed some fluctuation (91.43%, 91.34%, and 91.73%). When all three experts were combined (Expert 1 + Expert 2 + Expert 3), mIoU reached 89.10% and 91.69% at the early and compact stages, respectively, demonstrating the integration potential but with modest gains. However, introducing the MOEGatingNetwork module resulted in a significant performance breakthrough. On the early heading dataset, cabbage leaf IoU increased by 1.07% to 75.96% (Table 6, row 9), resolving conflicts among experts. On the compact heading dataset, cabbage head IoU improved by 0.37% to 86.11% (Table 7, row 9), with mIoU reaching 91.88% (+0.19%), achieved with zero increase in FLOPs (steady at 165.72 G) and only a negligible parameter increase of 0.08 M (<0.4%). The MOEGatingNetwork module dynamically allocates weights to select the optimal experts while suppressing ineffective interactions (such as the accuracy fluctuations seen in compact stage dual-expert combinations), enabling a synergistic effect where the whole exceeds the sum of its parts (early stage mIoU increased by 1.45% over baseline to 89.51%), thus realizing efficient fusion at no additional computational cost.

These ablation studies robustly validate the effectiveness and complementarity of the proposed heterogeneous expert collaboration architecture and its core modules in enhancing cabbage segmentation performance. The results demonstrate that the MSCross-DualAttention module effectively focuses on key regions, boosting model accuracy, while the gating fusion module (MOEGatingNetwork) plays a pivotal role in dynamically and adaptively integrating the strengths of different experts, significantly improving segmentation precision—especially for the critical cabbage head class—and ultimately achieving the optimal heterogeneous expert synergy embodied in U-MoEMamba. This provides an efficient and scalable solution for high-precision semantic segmentation in agricultural UAV remote sensing scenarios.

Figure 11 (early heading stage) and Figure 12 (compact heading stage) intuitively illustrate the visualization results of the ablation experiments. The comparison reveals that the multi-scale convolution expert (Expert 1) initially improves the main contours of the leaves (Figure 11, first row, third column), yet serrated boundaries appear around the leaf fold areas (highlighted by red dashed boxes), and the predictions deviate from the ground truth (compared to the Label column). Additionally, fragmentation is observed in the fruit texture segmentation (Figure 12, first row, third column). The incorporation of the attention mechanism (Expert 2) markedly enhances semantic coherence (Figure 11, first row, fourth column), producing more natural transitions in leaf fold textures and improving segmentation accuracy; however, sensitivity to dark regions of the fruit remains limited (red box in Figure 12, first row, fourth column). With the addition of the state-space model (Expert 3), long-range structural integrity is best preserved (e.g., main leaf veins in Figure 11, first row, fifth column), though some local details remain blurred (serrated leaf edges within red boxes in Figure 11, first row, fifth column). Progressive optimization validates the collaborative modules: dual-expert fusion (such as Expert 1 + Expert 2) smooths boundaries in fold areas (red box in Figure 11, second row, first column), but overlapping leaves exhibit adhesion and missegmentation (yellow region infiltration in Figure 11, second row, first column).

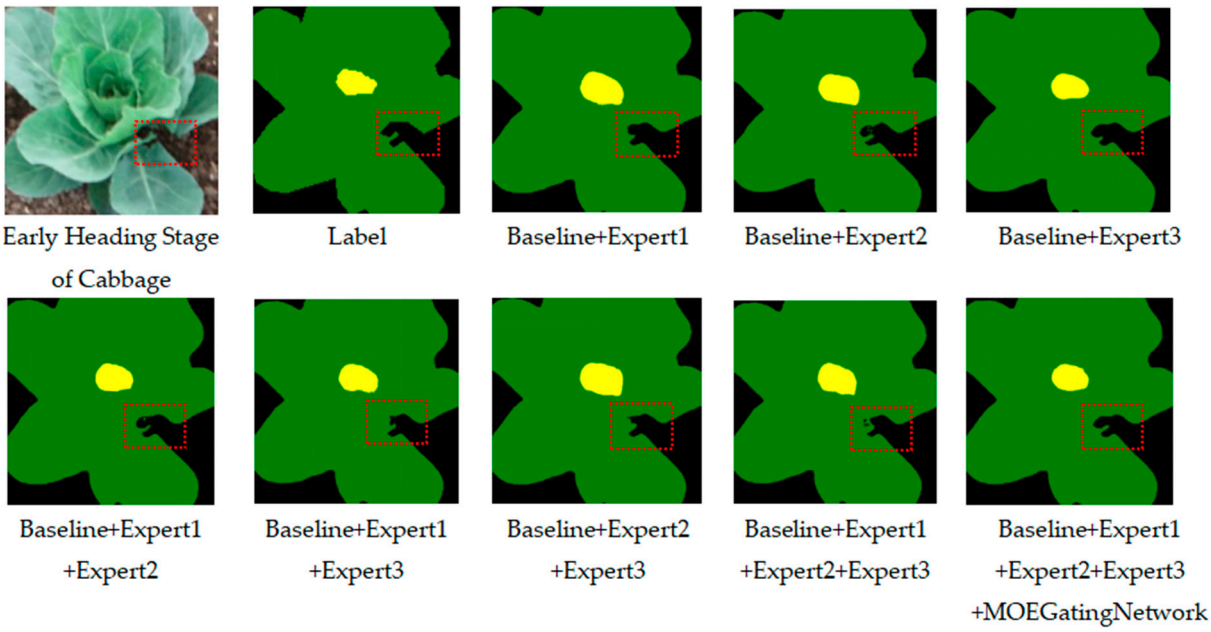


Figure 11. Visualization of ablation study results on the early heading stage cabbage dataset. Expert 1: Multi-scale Conv; Expert 2: MSCrossDualAttention; Expert 3: VSSBlock. The red dotted boxes highlight the visual differences among the results generated by different models.

When all three experts are integrated without gating fusion, residual noise appears in leaf gaps (black holes within red boxes in Figure 11, second row, fourth column), and details in dark fruit regions are lost (Figure 12, second row, fourth column). Upon introducing the gating mechanism (MOEGatingNetwork), the complete model U-MoEMamba precisely separates overlapping leaves in the early heading stage (no infiltration in yellow regions, second row, final column of Figure 11), while restoring subpixel continuity in leaf folds (continuous grooves within red boxes). During the compact heading stage, the dark fruit areas are completely segmented (red box in final column of Figure 12, second row), eliminating fragmentation artifacts observed in dual-expert models (compared to Figure 12, second row, first column). Boundary optimization is also achieved, with a significant reduction in serrated fruit edges (the second row, final column of Figure 12 shows markedly

smoother contours than the second row, third column), enabling pixel-level adherence to ground truth labels.

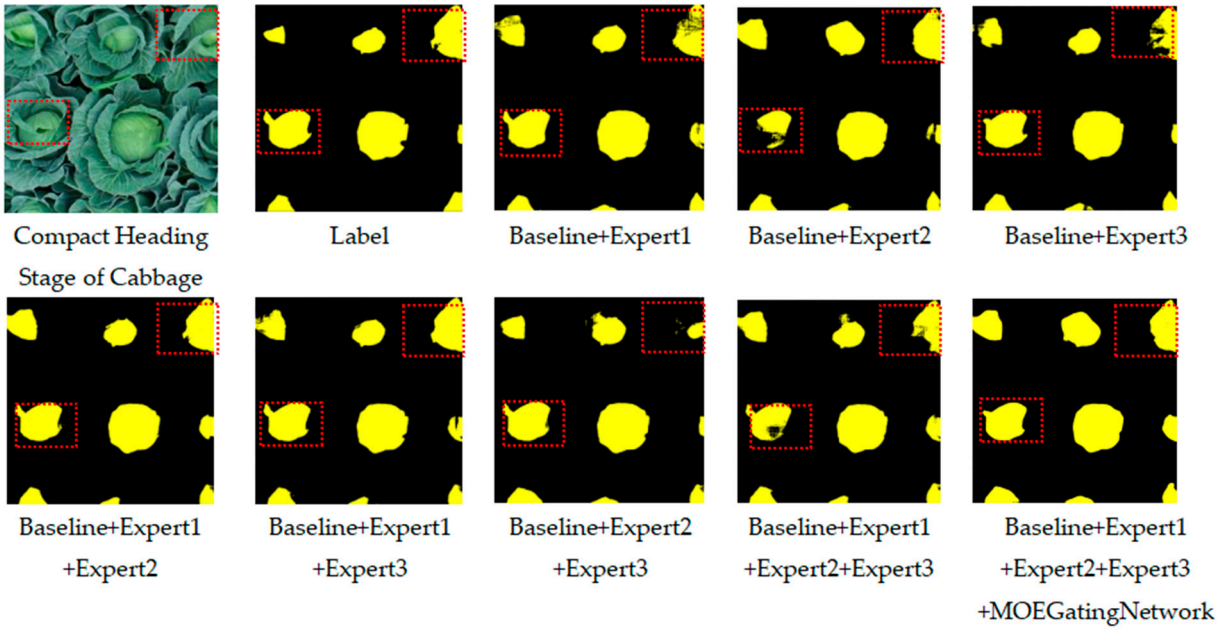


Figure 12. Visualization of ablation study results on the compact heading stage cabbage dataset. Expert 1: Multi-scale Conv; Expert 2: MSCrossDualAttention; Expert 3: VSSBlock. The red dotted boxes highlight the visual differences among the results generated by different models.

4. Discussion

UAV-based low-altitude remote sensing for cabbage head segmentation in complex field environments faces multiple challenges, including irregular geometric shapes of cabbage heads causing blurred semantic boundaries, overlapping leaves under dense planting conditions, and feature drift induced by dynamic lighting, all of which contribute to reduced segmentation accuracy. This study proposes the U-MoEMamba framework, which deeply integrates state-space models (SSMs) with a dynamic Mixture-of-Experts (MoE) mechanism to effectively mitigate segmentation difficulties arising from varying illumination, irregular shapes, leaf occlusion, and similar target textures in complex scenarios, thereby achieving pixel-level precise segmentation. Methodologically, the framework introduces novel components, including the MSCrossDualAttention module, the MambaMoEFusion module, and the MoEGatingNetwork module. By establishing multiple expert branches and leveraging Mamba's SSM mechanism, it dynamically fuses features and models long-range dependencies to meet multi-scale feature perception demands, substantially enhancing segmentation accuracy.

Experiments conducted on the cabbage dataset collected from Dashu Community, Xiushan Street, Tonghai County, Yunnan Province demonstrate that U-MoEMamba outperforms RS3Mamba by achieving boundary F1 score improvements of 1.03% and 0.27%, respectively, while reducing the parameter count by half. Additionally, mIoU is increased by 1.45% and 1.36%. Compared with multiple mainstream Mamba-based methods, U-MoEMamba achieves the best performance on nearly all evaluation metrics.Visual comparisons of segmentation results reveal that models such as SegMamba and CNN-Mamba UNet tend to suffer from class boundary confusion, manifesting as blurred segmentation edges, target merging, or omission of small objects. In contrast, the proposed U-MoEMamba significantly outperforms these models, excelling in fine-grained segmentation, class discrimination, and structural integrity. Ablation studies confirm that the MSCrossDualAtten-

tion module (Expert 2) effectively focuses on critical regions to enhance model accuracy, while the MOEGatingNetwork module plays a pivotal role within the three-expert collaborative architecture by dynamically and adaptively integrating the strengths of different experts, substantially improving segmentation precision—particularly for key targets such as cabbage heads—and ultimately achieving the optimal performance of the heterogeneous expert collaborative framework embodied in U-MoEMamba. Although the proposed method demonstrates promising performance in fine-grained cabbage segmentation, several limitations remain and merit further exploration:

- Sensitivity to Extreme Weather Conditions: The current model was primarily trained and evaluated under common weather scenarios such as sunny and cloudy conditions. Its performance in more challenging environments, such as heavy rainfall and strong winds, has not yet been thoroughly examined, which may affect its robustness in real-world deployments.
- Dependence on High-Performance Hardware: Due to the integration of multiple expert modules, the model requires considerable computational resources, particularly during training. This limits its immediate applicability on edge devices, indicating a need for lightweight optimization in future work.
- Limited Generalizability to Other Crops: The current study focuses exclusively on cabbage segmentation, with limited applicability to other crops. The model's adaptability and effectiveness across different crop types with similar morphological characteristics (e.g., lettuce, romaine) remain to be validated.

In future research, we plan to: (1) build a more diverse dataset covering various climatic conditions and cabbage cultivars, (2) explore model compression and knowledge distillation techniques to enhance deployment efficiency on low-resource devices, and (3) extend our approach to other vegetable crops to evaluate the generality and scalability of the proposed architecture.

5. Conclusions

In this study, we propose a novel semantic segmentation framework named U-MoEMamba, which integrates a ResT encoder for extracting joint local-global features, a hybrid Mixture-of-Experts (MoE) system, and the Mamba module to construct a dynamically differentiable state-space model tailored for low-altitude UAV remote sensing. The proposed framework achieves fine-grained and high-precision segmentation under complex field conditions. The key conclusions and original contributions of this work are summarized as follows:

1. U-MoEMamba introduces a heterogeneous expert collaboration architecture, which comprises three complementary expert branches: a multi-scale convolutional expert, a global attention expert (MSCrossDualAttention), and a long-range dependency modeling expert (VSSBlock). These experts are adaptively fused at the pixel level through a gated fusion module (MambaMoEFusion), enabling optimal feature representation for different spatial regions and significantly enhancing segmentation robustness in complex agricultural scenes. The proposed MSCrossDualAttention module, as one of the expert branches, effectively captures global contextual dependencies in the image, thereby strengthening the model's semantic understanding of entire scenes.
2. Compared with mainstream Mamba-based segmentation models (e.g., EfficientPyramidMamba and SegMamba), U-MoEMamba demonstrates superior performance on datasets corresponding to two key growth stages of cabbage. Specifically, it achieves 89.51% mIoU and 97.85% overall accuracy (OA) on the early heading dataset, and 91.88% mIoU and 97.98% OA on the compact heading dataset, with notable improvements in missed and incorrect segmentation.

3. Trained and evaluated on UAV datasets collected under diverse real-world conditions—including overcast and sunny weather, varying degrees of leaf occlusion, and different planting densities—U-MoEMamba exhibits strong generalization and environmental adaptability. Its practical potential is significant for smart agriculture, particularly in tasks such as crop growth monitoring, heading stage assessment, and yield prediction. Furthermore, this work aligns with the broader trend of AI-driven agricultural automation and digital phenotyping, supporting labor-efficient, data-informed, and sustainable crop management. It offers a reliable technical foundation for advancing precision farming practices.

Author Contributions: Conceptualization, R.L.; methodology, R.L.; software, R.L.; validation, R.L., X.D., S.P. and F.C.; formal analysis, R.L.; investigation, R.L.; resources, R.L.; data curation, R.L.; writing—original draft preparation, R.L.; writing—review and editing, X.D. S.P. and F.C.; visualization, R.L.; supervision, X.D. S.P. and F.C.; project administration, R.L.; funding acquisition, X.D. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Yunnan Basic Research Program, Project Title: Remote sensing estimation of above-ground carbon sinks in vegetation of urban agglomerations in central Yunnan and its response to climate change and human activities (Grant No. 202401AT070103); the National Natural Science Foundation of China, Project Title: Dynamic evolution, driving mechanism and optimization simulation of production-life-ecological space conflict in Central Yunnan Urban Agglomeration (Grant No. 42261073), and Project Title: Dynamic coupling mechanism and simulation of optimization control of land use and ecosystem services in the process of rapid urbanization in central Yunnan (Grant No. 41971369).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Additional data used to support the results of this study are available from the corresponding author upon request.

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

References

1. Moreb, N.; Murphy, A.; Jaiswal, S.; Jaiswal, A.K. Cabbage. In *Nutritional Composition and Antioxidant Properties of Fruits and Vegetables*; Academic Press: Cambridge, MA, USA, 2020; pp. 33–54.
2. Vrchota, J.; Pech, M.; Švepešová, I. Precision agriculture technologies for crop and livestock production in the Czech Republic. *Agriculture* **2022**, *12*, 1080. [CrossRef]
3. Zhang, Z.; Lin, M.; Li, D.; Wu, R.; Lin, R.; Yang, C. An AUV-enabled dockable platform for long-term dynamic and static monitoring of marine pastures. *IEEE J. Ocean. Eng.* **2025**, *50*, 276–293. [CrossRef]
4. Puppala, H.; Peddinti, P.R.; Tamvada, J.P.; Ahuja, J.; Kim, B. Barriers to the adoption of new technologies in rural areas: The case of unmanned aerial vehicles for precision agriculture in India. *Technol. Soc.* **2023**, *74*, 102335. [CrossRef]
5. Huang, H.; Deng, J.; Lan, Y.; Yang, A.; Deng, X.; Wen, S.; Zhang, H.; Zhang, Y. Accurate weed mapping and prescription map generation based on fully convolutional networks using UAV imagery. *Sensors* **2018**, *18*, 3299. [CrossRef]
6. Sahin, H.M.; Miftahushudur, T.; Grieve, B.; Yin, H. Segmentation of weeds and crops using multispectral imaging and CRF-enhanced U-Net. *Comput. Electron. Agric.* **2023**, *211*, 107956. [CrossRef]
7. Kang, S.; Li, D.; Li, B.; Zhu, J.; Long, S.; Wang, J. Maturity identification and category determination method of broccoli based on semantic segmentation models. *Comput. Electron. Agric.* **2024**, *217*, 108633. [CrossRef]
8. Crespo, A.; Moncada, C.; Crespo, F.; Morocho-Cayamcela, M.E. An efficient strawberry segmentation model based on Mask R-CNN and TensorRT. *Artif. Intell. Agric.* **2025**, *15*, 327–337. [CrossRef]
9. Xiang, D.; He, D.; Sun, H.; Gao, P.; Zhang, J.; Ling, J. HCMPE-Net: An unsupervised network for underwater image restoration with multi-parameter estimation based on homology constraint. *Opt. Laser Technol.* **2025**, *186*, 112616. [CrossRef]
10. Yu, Y.; Zhu, F.; Qian, J.; Fujita, H.; Yu, J.; Zeng, K.; Chen, E. CrowdFPN: Crowd counting via scale-enhanced and location-aware feature pyramid network. *Appl. Intell.* **2025**, *55*, 359. [CrossRef]
11. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]

12. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
13. Xiang, P.; Pan, F.; Liu, T.; Zhao, X.; Hu, M.; He, D.; Zhang, B. MixSegNext: A CNN-Transformer hybrid model for semantic segmentation and picking point localization algorithm of Sichuan pepper in natural environments. *Comput. Electron. Agric.* **2025**, *237*, 110564. [CrossRef]
14. Cai, Y.; Zeng, F.; Xiao, J.; Ai, W.; Kang, G.; Lin, Y.; Cai, Z.; Shi, H.; Zhong, S.; Yue, X. Attention-aided semantic segmentation network for weed identification in pineapple field. *Comput. Electron. Agric.* **2023**, *210*, 107881. [CrossRef]
15. Liu, L.B.; Zhao, C.J.; Wu, H.R.; Gao, R.H. Image Reduction Method for Rice Leaf Disease Based on Visual Attention Model. *Appl. Mech. Mater.* **2012**, *220*, 1393–1397. [CrossRef]
16. Meng, F.; Li, J.; Zhang, Y.; Qi, S.; Tang, Y. Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks. *Comput. Electron. Agric.* **2023**, *214*, 108298. [CrossRef]
17. Wang, J.; Zhang, Z.; Luo, L.; Wei, H.; Wang, W.; Chen, M.; Luo, S. DualSeg: Fusing transformer and CNN structure for image segmentation in complex vineyard environment. *Comput. Electron. Agric.* **2023**, *206*, 107682. [CrossRef]
18. Guo, Z.; Cai, D.; Jin, Z.; Xu, T.; Yu, F. Research on unmanned aerial vehicle (UAV) rice field weed sensing image segmentation method based on CNN-transformer. *Comput. Electron. Agric.* **2025**, *229*, 109719. [CrossRef]
19. Jeon, Y.-J.; Hong, M.J.; Ko, C.S.; Park, S.J.; Lee, H.; Lee, W.-G.; Jung, D.-H. A hybrid CNN-Transformer model for identification of wheat varieties and growth stages using high-throughput phenotyping. *Comput. Electron. Agric.* **2025**, *230*, 109882. [CrossRef]
20. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
21. Xu, R.; Yang, S.; Wang, Y.; Cai, Y.; Du, B.; Chen, H. Visual mamba: A survey and new outlooks. *arXiv* **2024**, arXiv:2404.18861.
22. Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; Wang, Z.; Ye, Z. A survey on visual mamba. *Appl. Sci.* **2024**, *14*, 5683. [CrossRef]
23. Shi, D.; Li, C.; Shi, H.; Liang, L.; Liu, H.; Diao, M. A Hierarchical Feature-Aware Model for Accurate Tomato Blight Disease Spot Detection: Unet with Vision Mamba and ConvNeXt Perspective. *Agronomy* **2024**, *14*, 2227. [CrossRef]
24. Zhao, F.; He, Y.; Song, J.; Wang, J.; Xi, D.; Shao, X.; Wu, Q.; Liu, Y.; Chen, Y.; Zhang, G. Smart UAV-assisted blueberry maturity monitoring with Mamba-based computer vision. *Precis. Agric.* **2025**, *26*, 56. [CrossRef]
25. Zhang, X.; Mu, W. GMamba: State space model with convolution for Grape leaf disease segmentation. *Comput. Electron. Agric.* **2024**, *225*, 109290. [CrossRef]
26. Gao, X.; Wang, G.; Zhou, Z.; Li, J.; Song, K.; Qi, J. Performance and speed optimization of DLV3-CRSNet for semantic segmentation of Chinese cabbage (*Brassica pekinensis* Rupr.) and weeds. *Crop Prot.* **2025**, *195*, 107236. [CrossRef]
27. Kong, X.; Li, A.; Liu, T.; Han, K.; Jin, X.; Chen, X.; Yu, J. Lightweight cabbage segmentation network and improved weed detection method. *Comput. Electron. Agric.* **2024**, *226*, 109403. [CrossRef]
28. Ma, Z.; Wang, G.; Yao, J.; Huang, D.; Tan, H.; Jia, H.; Zou, Z. An improved U-net model based on multi-scale input and attention mechanism: Application for recognition of Chinese cabbage and weed. *Sustainability* **2023**, *15*, 5764. [CrossRef]
29. Tian, Y.; Cao, X.; Zhang, T.; Wu, H.; Zhao, C.; Zhao, Y. CabbageNet: Deep Learning for High-Precision Cabbage Segmentation in Complex Settings for Autonomous Harvesting Robotics. *Sensors* **2024**, *24*, 8115. [CrossRef] [PubMed]
30. Guan, Y.; Cui, Z.; Zhou, W. Reconstruction in off-axis digital holography based on hybrid clustering and the fractional Fourier transform. *Opt. Laser Technol.* **2025**, *186*, 112622. [CrossRef]

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.