remote sensing

MDPI

Article

DBFormer: A Dual-Branch Adaptive Remote Sensing Image Resolution Fine-Grained Weed Segmentation Network

Xiangfei She [1,2], Zhankui Tang [1,2,*], Xin Pan [1,2], Jian Zhao [1,2] and Wenyu Liu [3]

[1] School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun 130012, China; rj_sxf@ccit.edu.cn (X.S.); panxin@ccit.edu.cn (X.P.); zhaojian@ccit.edu.cn (J.Z.)
[2] Jilin Provincial Key Laboratory of Changbai Historical Culture and VR Reconstruction Technology, Changchun 130012, China
[3] School of Foreign Languages and Cultures, Jilin University, Changchun 130012, China; liuwenyu@jlu.edu.cn
* Correspondence: tangzhankui@stu.ccit.edu.cn

Abstract

Remote sensing image segmentation holds significant application value in precision agriculture, environmental monitoring, and other fields. However, in the task of fine-grained segmentation of weeds and crops, traditional deep learning methods often fail to balance global semantic information with local detail features, resulting in over-segmentation or under-segmentation issues. To address this challenge, this paper proposes a segmentation model based on a dual-branch Transformer architecture—DBFormer—to enhance the accuracy of weed detection in remote sensing images. This approach integrates the following techniques: (1) a dynamic context aggregation branch (DCA-Branch) with adaptive downsampling attention to model long-range dependencies and suppress background noise, and (2) a local detail enhancement branch (LDE-Branch) leveraging depthwise-separable convolutions with residual refinement to preserve and sharpen small weed edges. An Edge-Aware Loss module further reinforces boundary clarity. On the Tobacco Dataset, DBFormer achieves an mIoU of 86.48%, outperforming the best baseline by 3.83%; on the Sunflower Dataset, it reaches 85.49% mIoU, a 4.43% absolute gain. These results demonstrate that our dual-branch synergy effectively resolves the global–local conflict, delivering superior accuracy and stability in the context of practical agricultural applications.

Keywords: precision agriculture; fine-grained weed segmentation; dual-branch Transformer architecture; dynamic context aggregation; local detail enhancement; edge-aware loss

check for updates

Academic Editor: Xinghua Li

Received: 17 April 2025
Revised: 18 June 2025
Accepted: 25 June 2025
Published: 26 June 2025

Citation: She, X.; Tang, Z.; Pan, X.; Zhao, J.; Liu, W. DBFormer: A Dual-Branch Adaptive Remote Sensing Image Resolution Fine-Grained Weed Segmentation Network. *Remote Sens.* **2025**, *17*, 2203. https://doi.org/10.3390/rs17132203

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

In recent years, with the rapid development of remote sensing technologies and the widespread availability of unmanned aerial vehicle (UAV) imagery, these technologies have played increasingly important roles in agricultural monitoring, environmental management, urban planning, and other fields. Remote sensing plays a vital role in agricultural and environmental management by providing timely, large-scale observations. For example, vegetation indices derived from satellite imagery can monitor crop vigor and detect disease or pest outbreaks, enabling farmers to optimize planting strategies for improved yield and quality [1]. It also supports land-use classification to guide resource planning and sustainable production layouts [2], assesses water quality by estimating chlorophyll-a concentrations for eutrophication monitoring [3], and tracks atmospheric pollutants such as $PM_{2.5}$ and PM_{10} in real time to inform air pollution control measures [4]. In urban planning, remote sensing imagery is also pivotal; it facilitates land-use classification and

urban expansion monitoring, which helps urban planners understand changes in land use and devise rational urban development strategies [5]. Additionally, by analyzing the spatial distribution of urban surface temperatures, remote sensing data provide scientific support for monitoring the urban heat-island effect and for the development of strategies to mitigate it [6].

Precision agriculture is an agricultural management concept based on the observation, measurement, and response techniques associated with the within-field variability of crops [7]. One of its primary objectives is to minimize environmental impact by reducing reliance on chemical inputs such as herbicides and pesticides. Owing to its wide coverage, real-time capabilities, and multispectral features, remote sensing imagery has become an essential data source in precision agriculture for monitoring crop growth and pest or disease occurrences. For example, multispectral remote sensing data can be used to analyze crop chlorophyll content and the normalized difference vegetation index (NDVI) to assess crop health in real time, thereby guiding precise fertilization [8]. High-resolution remote sensing imagery can also be employed to detect the spatial distribution of field pests and diseases, enabling early warning and targeted interventions, which reduces the dependency on chemical control methods [9]. Furthermore, multi-temporal remote sensing monitoring of soil moisture and crop water stress can optimize irrigation strategies, thus reducing water waste and alleviating environmental burdens [10].

Weed detection is particularly critical in agricultural management because weeds compete with crops for growth resources and can serve as vectors for pest and disease propagation, ultimately affecting crop yield and quality. For instance, in cornfields, high-resolution remote sensing imagery can be used to precisely distinguish between weed and crop regions, thereby guiding localized herbicide application to reduce competition for water and nutrients and enhance crop yield and quality [11]. Similarly, in wheat fields, the use of multispectral and hyperspectral data to monitor weed distribution facilitates the early identification of areas that may act as sources of pest and disease outbreaks, allowing for the implementation of targeted control strategies to safeguard crop health [12]. Recent studies have also demonstrated the effectiveness of integrating spatial–spectral features for fine-grained weed recognition in hyperspectral imagery. Some examples include a lightweight multi-scale spatio-spectral attention module used to achieve accurate weed identification and a structure-preserving band selection method used to retain informative bands with minimal redundancy [13,14]. These advances indicate the growing potential of hyperspectral imaging in agricultural weed monitoring, offering a more comprehensive understanding of weed–crop dynamics and paving the way for the development of autonomous weed-detection systems. Thus, efficient detection of weeds is essential for crop pest and disease management and represents a critical first step toward developing autonomous weed-detection systems.

Prior to deep learning, weed segmentation relied on handcrafted algorithms combining spectral thresholds and classical classifiers. Methods such as Otsu thresholding on NDVI and morphological filtering separated vegetation from soil [15–18], while machine learning models—e.g., SVMs applied to hyperspectral corn imagery and SVM + LDA for shape-based weed discrimination—leveraged hand-crafted features [19–21]. However, these approaches suffer from sensitivity to illumination changes and complex field backgrounds, and their manual feature engineering is labor-intensive and lacks generalizability.

Traditional machine-learning frameworks struggle to build discriminative features capable of distinguishing crops from weeds, due to the similar morphologies and spectral signatures of the plants. Deep learning, especially end-to-end CNNs, overcomes this by automatically learning hierarchical, multi-scale representations. While object detection approaches (e.g., bounding-box classifiers) facilitate coarse localization, they often miss fine

edge details and cannot resolve overlapping vegetation [22–24]. In comparison, semantic segmentation methods classify every pixel in an image, enabling more detailed target delineation and effectively capturing subtle differences between crops and weeds within complex backgrounds, thereby providing data for precision agricultural management which is higher-resolution and more reliable.

The most common semantic segmentation-based weed mapping methods utilize CNNs and Transformer architectures. Pixel-wise classification has been applied to distinguish weeds and crops in field imagery [25,26]. Traditional U-Net, with its effective multi-scale feature fusion, remains a popular baseline for agricultural weed detection [27]. DeepLabV3+ further expands the receptive field via atrous convolutions and integrates multi-level features for robust segmentation under complex backgrounds [28]. TransUNet combines Transformer encoders for global dependency modeling with U-Net decoders for local feature recovery, improving fine-grained weed mapping [29]. More recent lightweight architectures such as SegFormer capture both global context and local details efficiently [30], and CSSwin-Unet—an adaptation of Swin-Unet for high-resolution crop and field segmentation—demonstrates further gains [31]. AgriFM, a multi-source foundation model, extends this line of development by unifying spatiotemporal crop mapping across varied data modalities [32]. However, remote sensing images often suffer uneven resolution, dramatic scale variations, and rich fine-scale details, making accurate weed delineation challenging. In particular, the subtle differences in texture and morphology between weeds and their surrounding environments under complex backgrounds render fine-grained segmentation a challenging task.

To address the inherent conflict between global context modeling and local detail extraction in the fine-grained weed segmentation of remote sensing imagery, this paper innovatively proposes a dual-branch adaptive resolution fine-grained weed segmentation network (DBFormer). Built upon a Transformer framework [33], DBFormer features a unique dual-branch structure in its decoder stage, wherein each branch plays an indispensable role. The first branch, termed the dynamic context aggregation branch (DCA-Branch), serves as the primary backbone of the model. It employs an adaptive downsampling strategy to model multi-scale features and integrates attention mechanisms to enhance the expression of global contextual information, thereby enabling a comprehensive representation of the differences between crops and weeds in complex field environments. The second branch, known as the local detail enhancement branch (LDE-Branch), functions as an auxiliary branch that focuses on local convolutional transformations and the reinforcement of detailed features. This branch is particularly designed to refine and augment local features, thereby reducing the confusion between crops and weeds at their boundaries and enhancing the spatial consistency of segmentation results. The synergistic operation of these two branches allows DBFormer to retain global structural information while precisely capturing local details, ultimately improving the accuracy and robustness of weed segmentation. In experiments, our method was compared with five existing approaches, and the results demonstrate a significant improvement in overall accuracy and boundary semantic segmentation performance. The remainder of this paper is organized as follows: Section 2 describes the sources and preprocessing of the experimental data; Section 3 presents the overall structure of the proposed model and its constituent modules; Section 4 reports comparative results with existing methods; Section 5 provides an in-depth discussion of the research outcomes; and Section 6 concludes the paper.

2. Materials and Data

2.1. Data Sources

The experiments in this study utilized two datasets: the Tobacco Dataset and the Sunflower Dataset.

The Tobacco Dataset is a newly acquired set of tobacco weed images captured using a Mavic Mini UAV. The dataset was collected from eight tobacco fields in Mardan, Khyber Pakhtunkhwa, Pakistan. Data were captured at different growth stages, with the crop age ranging approximately from 15 to 40 days. Images were acquired at a resolution of 1920 × 1080 pixels, with an average altitude of 4 m and a ground sampling distance of 0.1 cm/pixel. Manual annotations were performed for background, crops, and weeds, with label values of 0, 1, and 2 assigned, respectively [34].

The Sunflower Dataset was obtained at a sunflower farm in Jesi, Italy, using a customized agricultural robot. This dataset was recorded during the spring season over a one-month period, spanning from the seedling stage to the end of the period during which chemical treatments remain effective. Images were captured using a 4-channel (RGB + NIR) JAI AD-13 camera mounted on the robot in a downward-facing configuration. The Sunflower Dataset comprises 500 images providing both RGB and NIR data, with pixel-level annotations for three classes: crops, weeds, and soil [35].

2.2. Dataset Preparation

For the Tobacco Dataset, all original images were cropped into non-overlapping patches of 480 × 352 pixels, yielding a total of 2520 image pairs with corresponding annotations. Using a fixed random seed (42), we then performed stratified random sampling to select 400 patches such that each contained between 10% and 50% weed pixels, covering diverse weed–crop density scenarios. These 400 pairs were partitioned into a training set of 320 pairs (80%), a validation set of 40 pairs (10%), and a test set of 40 pairs (10%).

For the Sunflower Dataset, given the original image size of 1296 × 964 pixels and system memory constraints, a non-overlapping cropping strategy was employed in a left-to-right and top-to-bottom order. Each original image was segmented into 4 non-overlapping patches of 640 × 480 pixels, as illustrated in Figure 1; the corresponding annotation images were cropped simultaneously. Again using seed 42 and stratified sampling on weed–crop ratio (10–50%), we randomly selected 400 patches. These were split into training (320 pairs, 80%), validation (40 pairs, 10%), and test (40 pairs, 10%) sets.

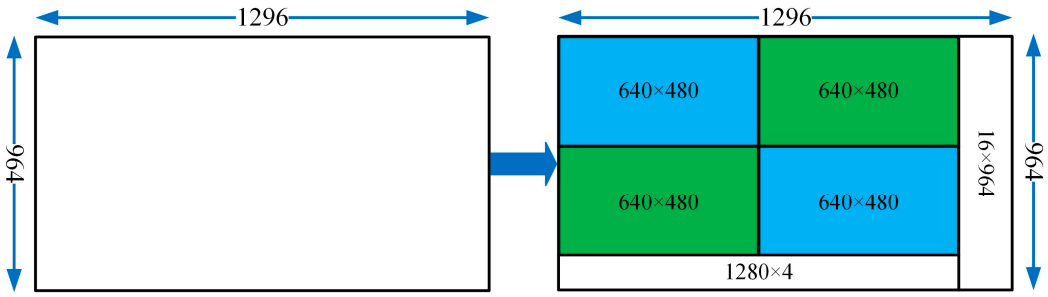


Figure 1. The procedure for cropping each original image from the Sunflower Dataset into four non-overlapping patches of 640 × 480 pixels, using code.

3. Methods

Aiming to achieve precise fine-grained weed segmentation in remote sensing imagery, DBFormer is designed to satisfy several key objectives simultaneously. Firstly, it must model global contextual information. Since the distribution patterns of weeds are influenced by field conditions and crop growth stages, relying solely on local features makes it difficult

to accurately distinguish weeds from crops. Therefore, the model must be capable of extracting global information from large-scale images, encompassing the background environment, vegetation distribution patterns, and the relationships between multi-scale features, to enhance overall segmentation stability. Secondly, it is essential to preserve local detail features. Weeds in agricultural fields typically exhibit complex morphological characteristics—such as slender leaves and small vegetation clusters—that are prone to being lost during global feature extraction. Thus, while focusing on global information, the model must accurately capture local texture and edge details to ensure precise identification of small weed targets, thereby improving segmentation accuracy and robustness. Finally, maintaining the integrity and clarity of boundaries is crucial. Due to the morphological similarities between weeds and crops, combined with challenges such as illumination variations and shadow interference in remote sensing imagery, target boundaries may become blurred or fragmented. Consequently, the segmentation model must enhance its capability to extract boundary features; this results in segmentation outputs which are more complete and with clear boundaries, which in turn minimizes the confusion between crops and weeds and provides high-quality data support for precision agriculture management. Based on these objectives, we propose a novel approach—DBFormer—which integrates a dynamic context aggregation branch (DCA-Branch), a local detail enhancement branch (LDE-Branch), and a global–local feature fusion strategy. This design not only reinforces global feature modeling but also effectively preserves local details and boundary information. The following sections provide a detailed description of the design philosophy of DBFormer and its application in weed segmentation within remote sensing imagery.

3.1. Overall Framework of the Model

As shown in Figure 2, the overall architecture of DBFormer comprises two core components: an encoder and a decoder.

Input and Output: The model accepts a multi-channel remote sensing image block I_{image} as input and produces a segmentation result I_{result} after the processing through DBFormer.

1. Encoder $M_{Encoder}$

The decoder is responsible for gradually upsampling the encoded features $F_{Encoder}$ and fusing multi-level information through a dual-branch strategy, ultimately producing the segmentation result I_{result}. Its core modules include the following:

- **Feature Embedding and Hierarchical Processing:** Initially, the input image is divided into overlapping patches (overlap patch embedding) to preserve local continuity while reducing resolution and computational complexity. Subsequently, four cascaded Transformer blocks are employed, each using an overlap patch-merging mechanism to efficiently transmit information while reducing dimensionality.
- **Efficient Self-Attention (Efficient Self-Attn)**: The encoder incorporates a self-attention mechanism to effectively capture long-range dependencies, thereby enhancing the model's global contextual understanding of complex agricultural areas.
- **Mix–Feed Forward Network (Mix-FFN)**: Following each Transformer block, a mix-FFN is applied to perform nonlinear feature transformations, enriching the feature representation and improving the model's ability to discriminate field characteristics.
- **Feature Output**: After multiple layers of Transformer processing, the encoder generates a deep feature map $F_{Encoder}$, which is then passed to the decoder for further processing.

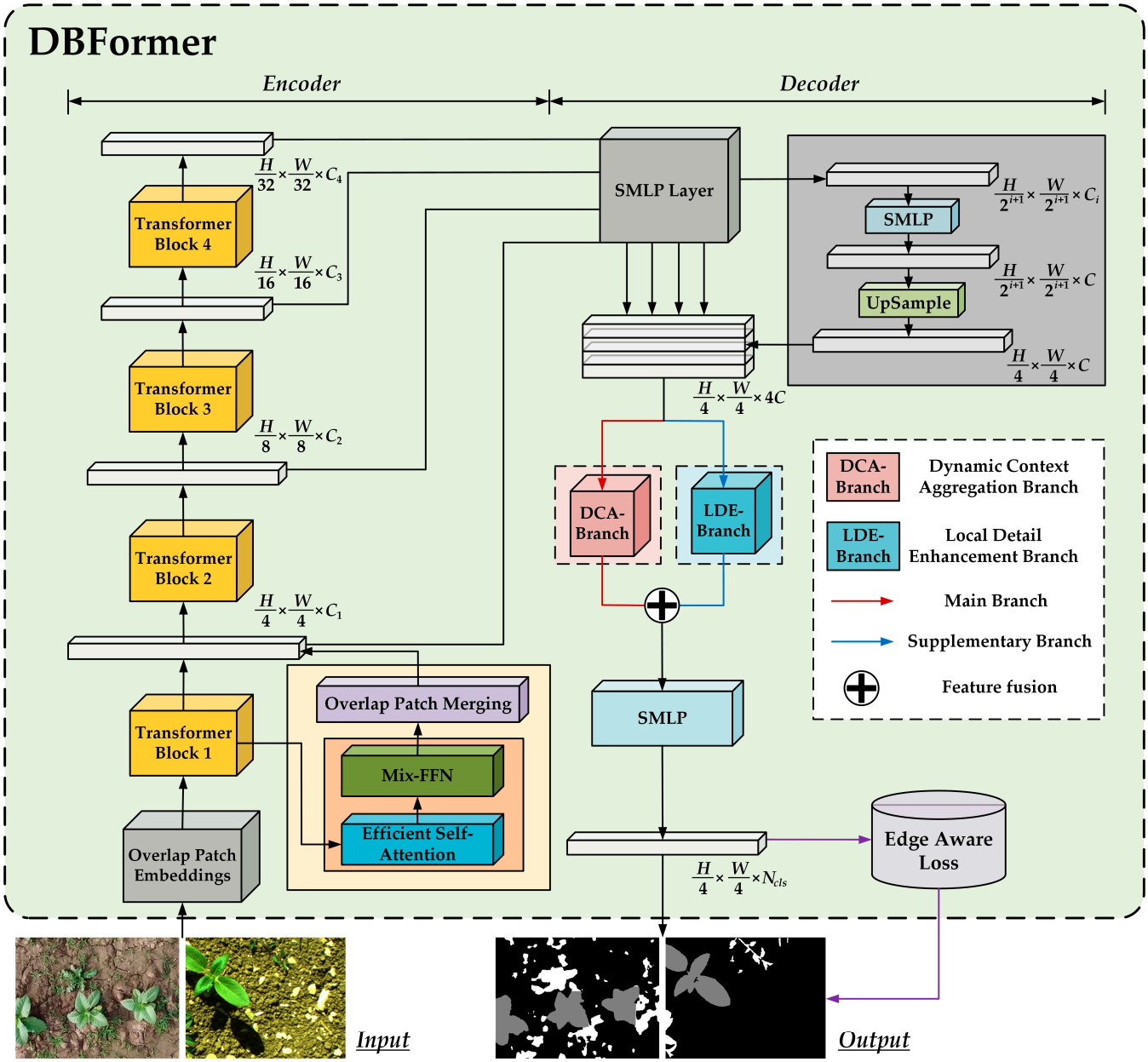


Figure 2. Overall architecture of DBFormer.

2. Decoder $M_{Decoder}$

The decoder is responsible for gradually upsampling the encoded features $F_{Encoder}$ and fusing multi-level information through a dual-branch strategy, ultimately producing the segmentation result I_{result}. Its core modules include the following:

- **SMLP Feature Transformation**: In the decoding stage, the SMLP structure is first applied to map the features, enhancing the compatibility among multi-scale features to ensure effective fusion across different levels.
- **Dynamic Context Aggregation Branch (DCA-Branch)**: This branch is primarily responsible for modeling global contextual information. By employing adaptive downsampling and attention mechanisms to perform weighted fusion on multi-scale features, it strengthens cross-scale information interaction, enabling a more accurate recognition of large-scale field regions.
- **Local Detail Enhancement Branch (LDE-Branch)**: Focused on extracting local details, this branch utilizes localized convolution operations and detail enhancement

strategies to strengthen the boundary features of small field regions, ensuring precise segmentation of minute weed targets.

- **Feature Fusion and Edge-Aware Loss**: At the end of the decoder, a feature fusion operation integrates the features extracted by the DCA-Branch and the LDE-Branch. An Edge-Aware Loss is then applied to reinforce field boundary information, thereby improving the model's capability to segment complex terrains and mixed-crop regions.
- **Final Output**: After multiple layers of SMLP processing, dual-branch fusion, and up-sampling operations, the decoder ultimately produces a high-precision segmentation result I_{result} with clear boundaries and coherent overall regions.

3.2. Model Encoder $M_{Encoder}$

The model encoder $M_{Encoder}$ employs a hierarchical Transformer-based architecture, which consists of overlap patch embeddings and four stages (Stage1, Stage2, Stage3, and Stage4), as illustrated in Figure 3. This encoder progressively extracts features from the input image while compressing its spatial dimensions, thereby generating encoded features that encapsulate multi-scale information.

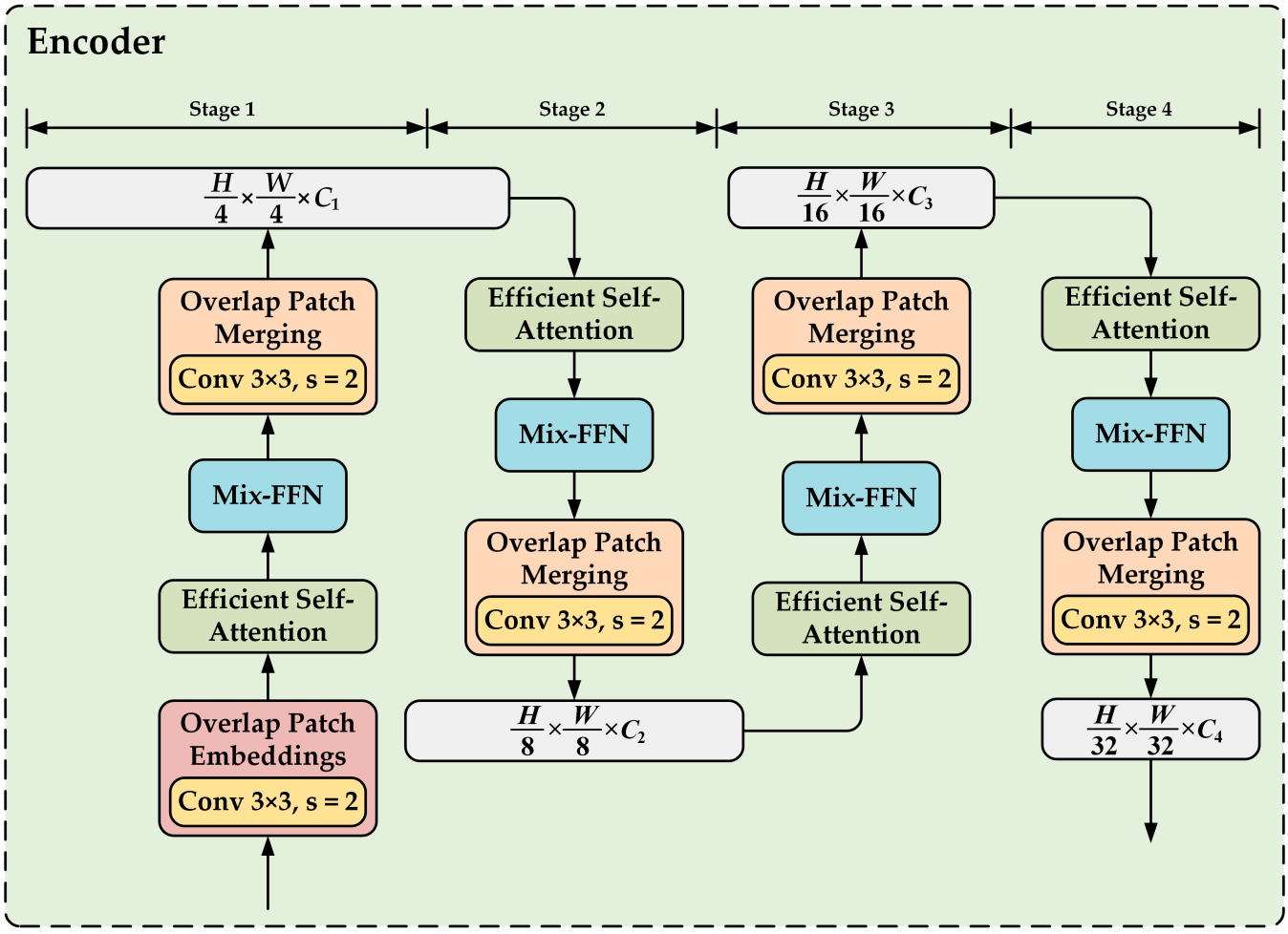


Figure 3. Illustration of the encoder architecture.

3.2.1. Overlap Patch Embeddings

At the input stage, a 3 × 3 convolutional kernel (with stride 2 and padding 1) is applied to the input image to obtain non-overlapping patch representations. Let the input image be $I_{image} \in \mathbb{R}^{H\times W\times 3}$. The patch embedding is computed as follows:

$$F_0 = \sigma\,(W_0 * I_{image} + b_0), \tag{1}$$

where W_0 denotes the convolution kernel weights, b_0 is the bias term, and $\sigma(\cdot)$ represents the ReLU activation function. This operation produces a feature map F_0 of dimensions $\frac{H}{2} \times \frac{W}{2} \times C_1$.

To further enrich local interactions, we follow with the MIX-FFN module (as in SegFormer) which interleaves point-wise and depth-wise convolutions:

$$U = \text{Conv}_{1\times1}(F_0; W_1, b_1), \tag{2}$$

$$V = \text{GELU}\,(\text{DWConv}_{3\times3}(U; W_{\text{dw}}, b_{\text{dw}})), \tag{3}$$

$$F_0' = F_0 + \text{Conv}_{1\times1}(V;\ W_2,\ b_2), \tag{4}$$

where W_1, b_1 project channels up by $4\times$; $\text{DWConv}_{3\times3}$ is a depth-wise convolution (kernel size = 3, stride = 1, padding = 1) with weights W_{dw} and bias b_{dw}; GELU denotes the Gaussian error linear unit activation; and W_2, b_2 project channels back to C_1.

The output F_0' then enters the subsequent Transformer blocks.

3.2.2. Stage 1: Initial Feature Extraction

In Stage 1, Transformer Block 1 is employed to extract features. Each Transformer block consists of an efficient self-attention (ESA) module and a mixed–feed forward network (Mix-FFN). First, the ESA module is applied to the feature map F_0, as described by the following formula:

$$\text{Attention}\ (Q,\ K,\ V) = \text{Softmax}\ (\frac{QK^T}{\sqrt{d_k}})V, \tag{5}$$

where $Q = W_Q F_0$, $K = W_K F_0$, and $V = W_V F_0$, with W_Q, W_K, and W_V being the linear transformation matrices, and d_k representing the attention dimension.

The output of the ESA is then passed through the Mix-FFN for nonlinear transformation, formulated as in the following:

$$F_1' = \text{Mix} - \text{FFN}\big(\text{ESA}\ (F_0')\big), \tag{6}$$

Simultaneously, overlap patch merging is applied to downsample the feature map and increase the channel dimension; this is computed as follows:

$$F_1 = W_1\ *\ F_1' + b_1, \tag{7}$$

This results in a feature map F_1 with dimensions $\frac{H}{4} \times \frac{W}{4} \times C_1$.

3.2.3. Stage 2: Multi-Scale Feature Extraction

In Stage 2, Transformer Block 2 further extracts features. The ESA and Mix-FFN computations are applied similarly (as in Equation (3)) to obtain an intermediate feature F_2', and patch merging is applied to further reduce the resolution, yielding the feature F_2. After Transformer Block 2, the resulting feature map F_2 has dimensions $\frac{H}{8} \times \frac{W}{8} \times C_2$.

3.2.4. Stage 3: High-Level Feature Extraction

Stage 3 utilizes Transformer Block 3 to extract higher-level semantic information. Similar to previous stages, ESA and Mix-FFN are applied to obtain F_3'; this is followed by patch merging to produce F_3. The final output of Stage 3 is a feature map F_3 with dimensions $\frac{H}{16} \times \frac{W}{16} \times C_3$.

3.2.5. Stage 4: Deep Feature Representation

In Stage 4, Transformer Block 4 is used for the final deep feature extraction. The ESA and Mix-FFN operations (as in Equation (3)) yield an intermediate feature F_4', and further

downsampling via patch merging produces F_4. The final feature map F_4 has dimensions $\frac{H}{32} \times \frac{W}{32} \times C_4$.

Ultimately, the features extracted from all stages (i.e., $\{F_1, F_2, F_3, F_4\}$) collectively form the deep feature map $F_{Encoder}$ with multi-scale information. This feature map is then fed into the SMLP layer for subsequent decoding.

3.3. Model Decoder $M_{Decoder}$

Corresponding to the encoder $M_{Encoder}$, the decoder receives multi-scale feature inputs from the encoder and progressively performs feature upsampling, fusion, and refinement. The decoder is also divided into several stages, including the SMLP stage, the dynamic context aggregation branch (DCA-Branch), the feature fusion with local detail enhancement branch (LDE-Branch), the Edge-Aware Loss, and the output stage, as shown in Figure 4. Its primary objective is to preserve global contextual information while accentuating local detail features, thereby achieving precise segmentation of fine-grained weeds in remote sensing imagery.

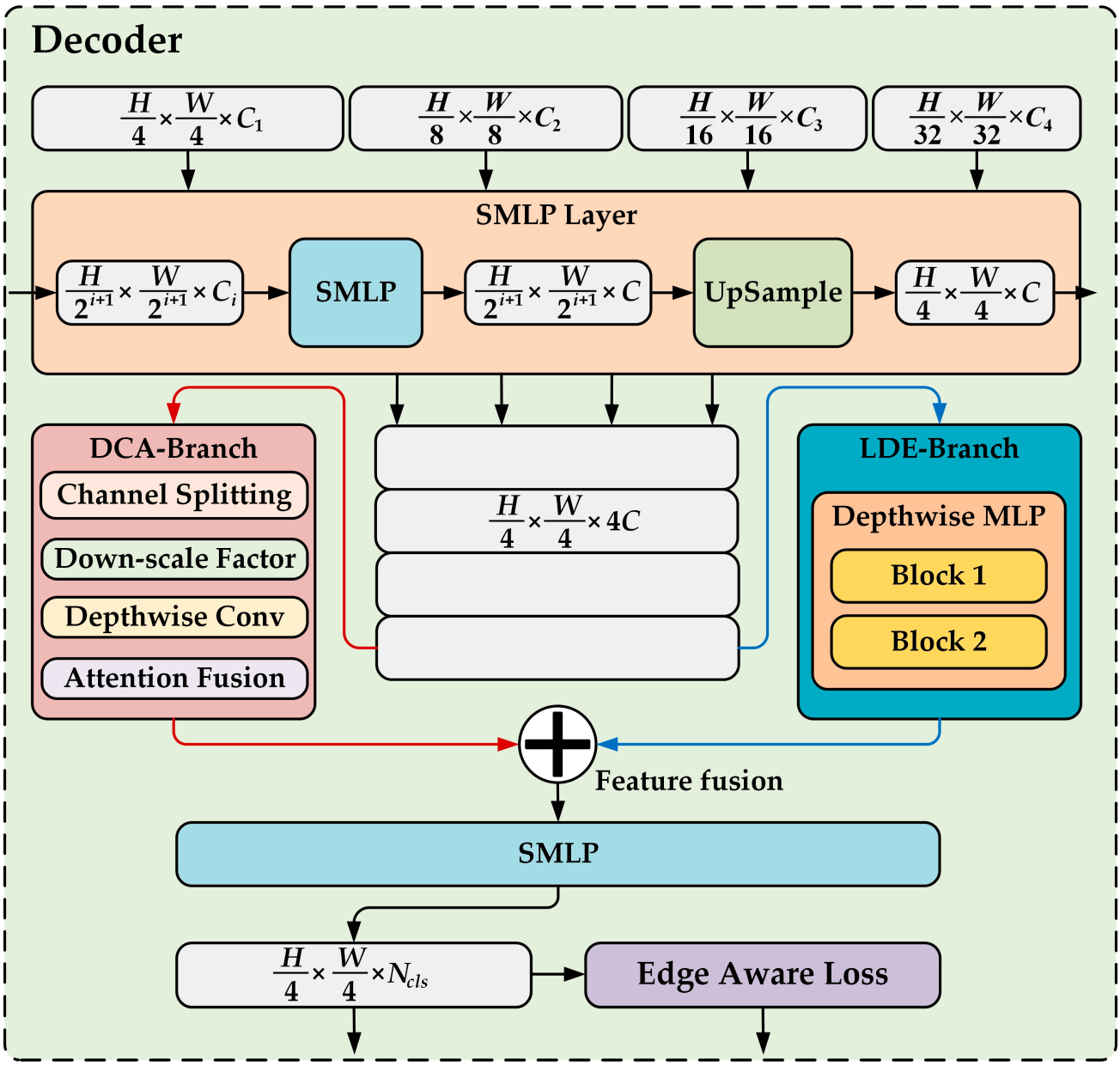


Figure 4. Illustration of the decoder architecture.

3.3.1. SMLP Stage

The feature maps F_1, F_2, F_3, and F_4 from the encoder are first combined to form the deep feature map $F_{Encoder}$. A lightweight SMLP module is then employed to align the channel dimensions of features at different resolutions, resulting in features of size $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C$. This alignment mitigates dimensional mismatches during subsequent fusion. Next, low-resolution features are upsampled to the target resolution of $\frac{H}{4} \times \frac{W}{4} \times C$ for consistency with high-resolution features. Finally, multi-scale features are concatenated along the channel dimension to yield the fused feature map F_{SMLP}, which serves as a unified input for the subsequent stages.

3.3.2. Dynamic Context Aggregation Branch (DCA-Branch)

The DCA-Branch obtains global contextual information by performing adaptive down-sampling and attention-based fusion, allowing flexible switching between different spatial scales to enhance the modeling of long-range dependencies. Its key steps include the following techniques.

1. Channel Splitting

The input feature F_{SMLP} is split into two parallel sub-branches, as shown in Equation (8):

$$LDE_{\mathbf{sub}}, DCA_{\mathbf{sub}} = \text{SplitConv}(F_{SMLP}), \tag{8}$$

where $LDE_{\mathbf{sub}}$ is used for local detail enhancement and $DCA_{\mathbf{sub}}$ for dynamic context aggregation.

2. Adaptive Down-sampling Factor Calculation

By analyzing the standard deviation σ of the current feature map, an adaptive down-sampling factor d is computed, enabling the network to adjust the scale of global context modeling based on scene or texture complexity, as given in Equation (9):

$$d = \max(1,\ min(base_scale,\ \lfloor base_scale \times \sigma / \beta \rfloor)), \tag{9}$$

where base_scale and β are tunable hyperparameters.

3. Down-sampling and Depthwise Convolution

The sub-branch $DCA_{\mathbf{sub}}$ is pooled to a size of $\frac{H}{d} \times \frac{W}{d}$ and then processed with two layers of 3×3 depthwise convolutions to extract global contextual features, as described in Equation (10):

$$F_{\mathbf{dca_down}} = \text{DWConv}(Pool(DCA_{sub})), \tag{10}$$

This operation enhances the capture of long-range pixel relationships while preserving global information.

4. Upsampling and Attention Fusion

The processed feature $F_{\mathbf{dca_down}}$ is upsampled back to the original size $H \times W$ and combined with the variance information of $DCA_{\mathbf{sub}}$ as input to an attention module, generating an attention map A, as given in Equation (11):

$$A = \text{Attention}\Big(F_{dca_up} \oplus Var(DCA_{sub})\Big), \tag{11}$$

Finally, the attention map is used to weight $DCA_{\mathbf{sub}}$, yielding the global feature $DCA_{\mathbf{fused}}$:

$$DCA_{\mathbf{fused}} = DCA_{\mathbf{sub}} \times A, \tag{12}$$

The DCA-Branch outputs $DCA_{\mathbf{fused}}$, which contains rich global contextual information and will be fused with the output from the local branch in subsequent stages.

3.3.3. Local Detail Enhancement Branch (LDE-Branch)

The LDE-Branch emphasizes the edge and texture features of fine-grained targets (e.g., small weeds) by preserving and enhancing local information through multiple convolutional layers and residual connections. It incorporates a key module called the Depthwise MLP (DMLP). In this module, the input feature $LDE_{\mathbf{sub}}$ is processed via two residual blocks:

1. The first block expands the channels using a 1×1 convolution, then applies a 3×3 depthwise convolution to extract local textures followed by an activation function;
2. The second block further enhances feature representation with another 3×3 depthwise convolution and lastly compresses the channels back to their original dimensions. The output is added to the input as a residual connection, as formulated in Equation (13):

$$F_{\mathbf{lde}} = \mathrm{DMLP}\,(LDE_{\mathbf{sub}}) = LDE_{\mathbf{sub}} + \Delta, \tag{13}$$

where Δ represents the feature increment obtained after the two depthwise convolutions.

3.3.4. Feature Fusion and Edge-Aware Loss

At the end of the decoder, the outputs of the DCA-Branch and the LDE-Branch are fused via an element-wise addition to yield the integrated feature $F_{\mathbf{fused}}$, as shown in Equation (14):

$$F_{\mathbf{fused}} = DCA_{\mathbf{fused}} + F_{\mathbf{lde}}, \tag{14}$$

In addition, a lightweight SMLP module is applied to align the channel dimensions of features from different resolutions. To further reinforce target boundaries, an Edge-Aware Loss is introduced to impose extra constraints on the gradients or edge regions of the prediction, thereby enhancing the clarity of fine-grained boundaries. Given the predicted mask M and ground-truth G, we extract ground-truth edges E_G via Sobel:

$$E_{\boldsymbol{G}} = |\nabla_{\mathrm{x}}\, G| + |\nabla_{\mathrm{y}}\, G|, \tag{15}$$

where ∇_{x} and ∇_{y} are the standard 3×3 Sobel kernels for horizontal and vertical gradient computation.

The edge loss is defined as the following:

$$L_{\mathrm{egde}} = \frac{1}{|E_G|}\sum\nolimits_{i \in E_G}(1 - IoU(M_i,\ G_i)), \tag{16}$$

The total training objective combines cross-entropy and edge-aware terms:

$$L = L_{\mathrm{CE}} + \lambda L_{\mathrm{egde}}, \tag{17}$$

where λ balances segmentation accuracy and boundary sharpness.

3.3.5. Output Stage

After multiple layers of SMLP processing, dual-branch fusion, and upsampling operations, the decoder maps the fused deep feature $F_{\mathbf{fused}}$ to the desired number of target classes (e.g., weed vs. non-weed) via a 1×1 convolution. The final output is the high-precision segmentation result $I_{\boldsymbol{result}}$, ensuring clear boundaries and coherent regions, which provides robust support for fine-grained weed monitoring in remote sensing imagery.

3.4. Experimental Environment

The experimental setup was configured as shown in Table 1:

Table 1. Experimental environment setup.

Name	Configuration
CPU	Intel Core(TM) i7-8700K 32 G, Intel Corporation, Santa Clara, CA, USA
GPU	NVIDIA GeForce RTX 2080 Ti, NVIDIA, Santa Clara, CA, USA
Operating System	Windows 10
Deep Learning Framework	Pytorch 1.12

The specific settings were as follows:

1. Hyperparameter Settings: The initial learning rate was set to 0.001 and adjusted using cosine annealing. The batch size was fixed at 4. The network was trained using the AdamW optimizer in conjunction with a cross-entropy loss to improve boundary sensitivity.
2. Training Settings: All experiments were conducted on a system equipped with an NVIDIA GeForce RTX 2080 Ti GPU, an Intel Core i7-8700K CPU, and 32 GB of RAM. The training process was carried out for 100 epochs with early stopping employed to prevent overfitting.

4. Results

4.1. Model Comparison

In this study, the primary models used for comparison include UNet, PSPNet, DeepLab V3+, SegFormer, PIDNet, and the proposed DBFormer. The UNet model, introduced in 2015, has become a benchmark for image segmentation due to its simplicity and effectiveness. PSPNet (Pyramid Scene Parsing Network), proposed in 2016, is a deep learning model for image segmentation mainly applied to scene parsing and semantic segmentation. DeepLab V3+ is a semantic segmentation model that captures multi-scale contextual information while preserving high resolution. SegFormer is a simple, efficient, and powerful semantic segmentation framework that combines Transformer with a lightweight multilayer perceptron decoder. PIDNet achieves real-time inference capabilities and low memory overhead while maintaining competitive accuracy, making it suitable for edge computing deployment in precision agriculture [36–40].

To ensure a comprehensive and reliable evaluation of the segmentation algorithms, we adopted widely recognized metrics in the field. These include a combination of pixel-based error metrics such as Precision, Pixel Accuracy, mIoU, and Recall [41–44]. The definitions of these pixel-based evaluation metrics are as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{18}$$

$$\text{PixelAccuracy} = \frac{\sum_i C_{ii}}{\sum_{i,j} c_{ij}}, \tag{19}$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{20}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{21}$$

where TP (true positives) denotes the number of samples correctly predicted to be positive; FP (false positives) denotes the number of samples incorrectly predicted to be positive; FN (false negatives) denotes the number of samples incorrectly predicted to be negative; C is the confusion matrix (also known as the error matrix), with C_{ii} representing the number of correctly classified pixels for class i, and $\sum_{i,j} C_{ij}$ the total number of pixels; N denotes

the number of classes; and TP_i, FP_i, and FN_i refer to the true positives, false positives, and false negatives for class i, respectively.

In addition, to address broader structural issues, object-based metrics, namely Over-segmentation (OS) and Under-segmentation (US), were incorporated [45–47]. Over-segmentation (OS) quantifies the proportion of extraneous pixels in the segmentation result relative to the sum of ground truth and extraneous pixels, as defined in Equation (22):

$$OS = \frac{O_s}{R_s + O_s}, \tag{22}$$

where O_s denotes the number of pixels in the segmentation result that should not be present (i.e., over-segmented pixels), and R_s denotes the number of pixels in the ground truth.

Under-segmentation (US) quantifies the proportion of omitted pixels in the segmentation result relative to the sum of ground truth and extraneous pixels, as defined in Equation (23):

$$US = \frac{U_s}{R_s + O_s} \tag{23}$$

where U_s represents the number of pixels omitted in the segmentation result (i.e., under-segmented pixels).

These metrics provide a comprehensive assessment of the segmentation algorithm's performance, particularly in terms of accuracy and completeness, when handling complex images.

4.2. Evaluation of Different Models

Initially, we trained the models—UNet, PSPNet, DeepLab V3+, SegFormer, PIDNet, and DBFormer—on both the Tobacco Dataset and Sunflower Dataset, using the same training set for 100 epochs. During training, the loss values for both datasets were recorded, as shown in Figure 5.

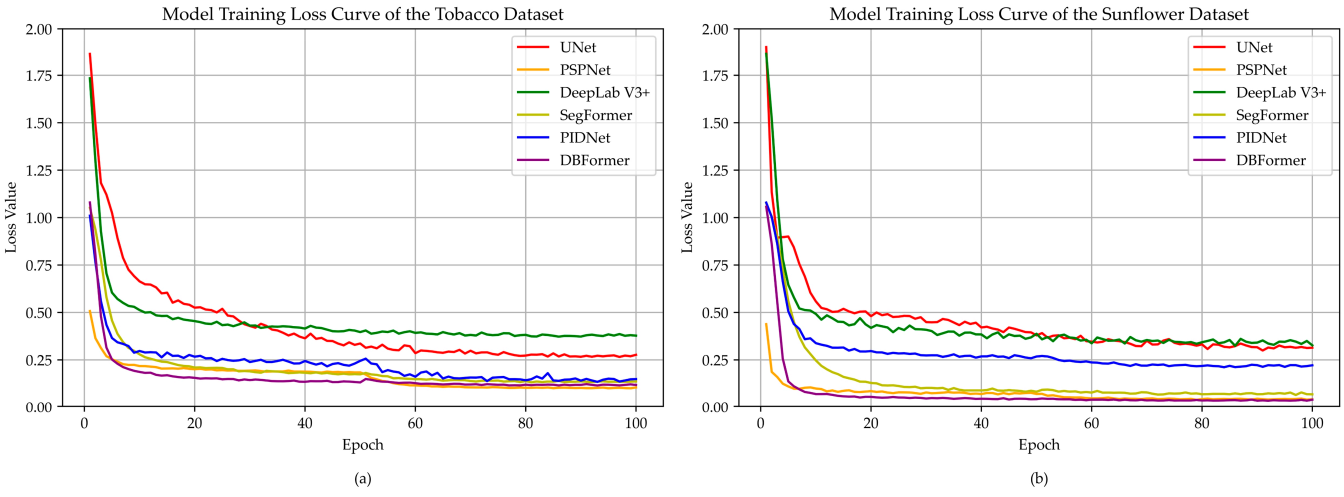


Figure 5. Training loss curves: (**a**) Model training loss curve of the Tobacco Dataset; (**b**)Model training loss curve of the Sunflower Dataset.

The results indicate that DBFormer exhibits faster convergence, stabilizing after approximately 50 epochs, whereas other models (e.g., UNet) exhibit oscillatory loss behavior due to inadequate boundary-feature learning. We attribute this to two key factors: (1) the DCA branch's adaptive downsampling reduces background noise and stabilizes early gradient updates, and (2) the LDE branch's residual refinement delivers stronger gradient signals at fine object boundaries. The synergy of these mechanisms accelerates convergence and guides the model toward a better local minimum.

Subsequently, the models were evaluated using five different groups of random seeds on a common validation set. The resulting Precision, Pixel Accuracy, mIoU, and Recall values are summarized in Table 2:

Table 2. Comparison of the model evaluation metrics (mean ± std over 5 runs).

Datasets	Model	Precision (%)	Pixel Accuracy (%)	mIoU (%)	Recall (%)
Tobacco Dataset	UNet [36]	87.72 ± 0.60	86.80 ± 0.55	78.46 ± 0.50	85.55 ± 0.60
	PSPNet [37]	88.05 ± 0.65	79.54 ± 0.60	72.78 ± 0.55	79.23 ± 0.60
	DeepLab V3+ [38]	83.36 ± 0.70	81.53 ± 0.65	72.24 ± 0.60	80.15 ± 0.65
	SegFormer [39]	87.62 ± 0.55	84.42 ± 0.50	76.57 ± 0.45	82.44 ± 0.50
	PIDNet [40]	89.92 ± 0.50	86.98 ± 0.45	82.65 ± 0.40	88.00 ± 0.50
	DBFormer (Ours)	92.89 ± 0.45	92.08 ± 0.50	86.48 ± 0.40	91.50 ± 0.50
Sunflower Dataset	UNet [36]	85.21 ± 0.65	81.17 ± 0.60	75.46 ± 0.55	80.12 ± 0.60
	PSPNet [37]	84.12 ± 0.70	76.44 ± 0.65	70.00 ± 0.60	75.80 ± 0.65
	DeepLab V3+ [38]	77.14 ± 0.75	80.82 ± 0.70	71.06 ± 0.65	79.50 ± 0.70
	SegFormer [39]	80.33 ± 0.70	70.44 ± 0.65	66.97 ± 0.60	68.97 ± 0.65
	PIDNet [40]	88.48 ± 0.55	83.92 ± 0.50	81.06 ± 0.45	83.30 ± 0.50
	DBFormer (Ours)	94.52 ± 0.35	88.73 ± 0.40	85.49 ± 0.45	88.00 ± 0.40

On the Tobacco Dataset, DBFormer achieves outstanding performance, with a Precision of 92.89 ± 0.45%, Pixel Accuracy of 92.08 ± 0.50%, mIoU of 86.48 ± 0.40%, and Recall of 91.50 ± 0.50%. These results clearly demonstrate its advantages in accurate prediction at the pixel level and its precise delineation of segmentation regions. Similarly, on the Sunflower Dataset, DBFormer exhibits excellent performance, with a Precision of 94.52 ± 0.35%, Pixel Accuracy of 88.73 ± 0.40%, mIoU of 85.49 ± 0.45%, and Recall of 88.00 ± 0.40%, reflecting high accuracy in pixel-level prediction and significant overlap between segmented and ground truth regions.

In contrast, UNet, PSPNet, DeepLab V3+, SegFormer, and PIDNet exhibit shortcomings in one or more metrics (such as Precision, Pixel Accuracy, mIoU, or Recall) on both datasets. For example, DeepLab V3+ has relatively low Precision, while UNet, PSPNet, SegFormer, and PIDNet show lower values in Pixel Accuracy, mIoU, and Recall. In comparison, DBFormer demonstrates superior performance across these metrics.

Overall, DBFormer outperforms the compared methods, especially in terms of Precision and mIoU, better meeting overall classification requirements and highlighting its potential for applications in precision agriculture.

4.3. Test Results of Different Models

We conducted a comparative evaluation of six models using the test sets from both the Tobacco Dataset and the Sunflower Dataset.

4.3.1. Tobacco Dataset

From the Tobacco Dataset test set, six groups of data were randomly selected for display, as shown in Figure 6.

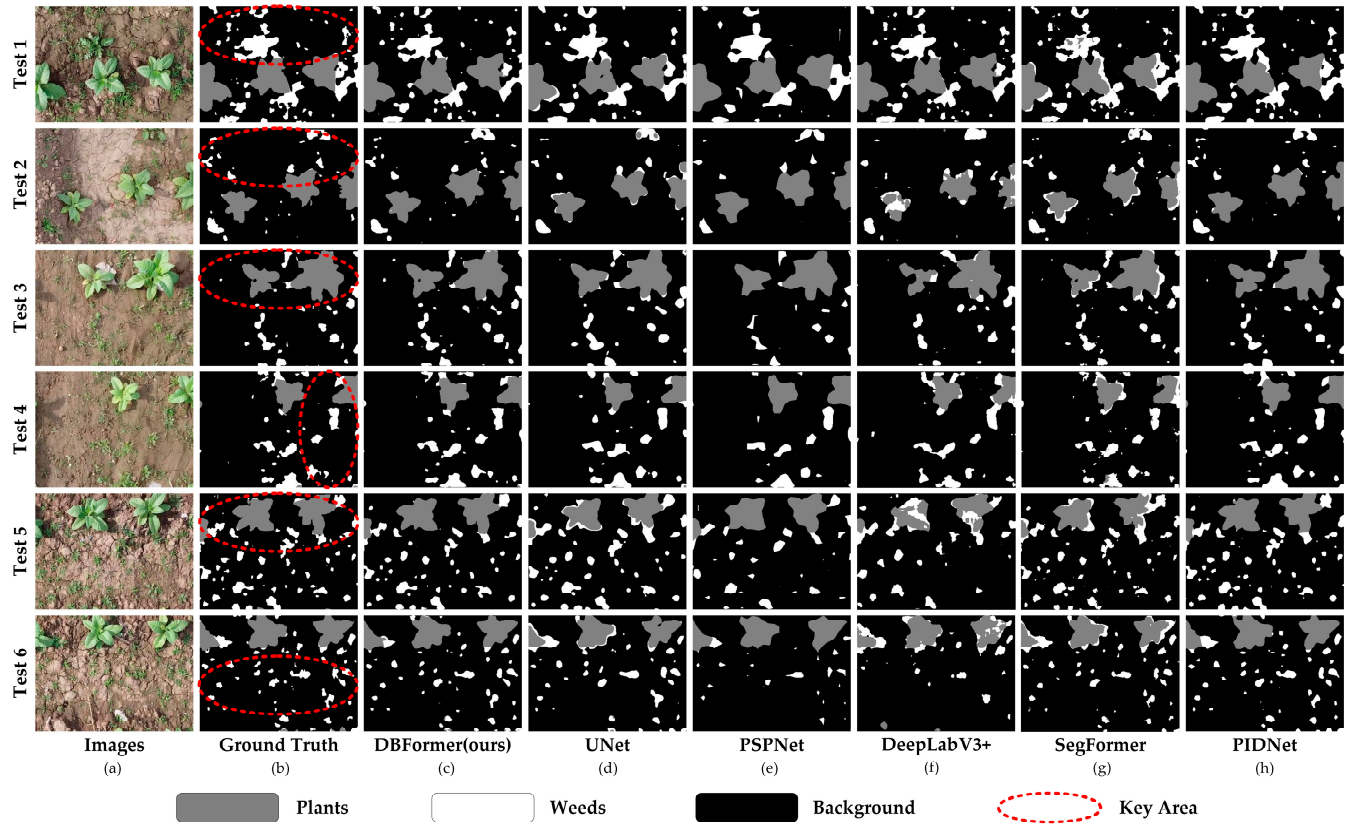
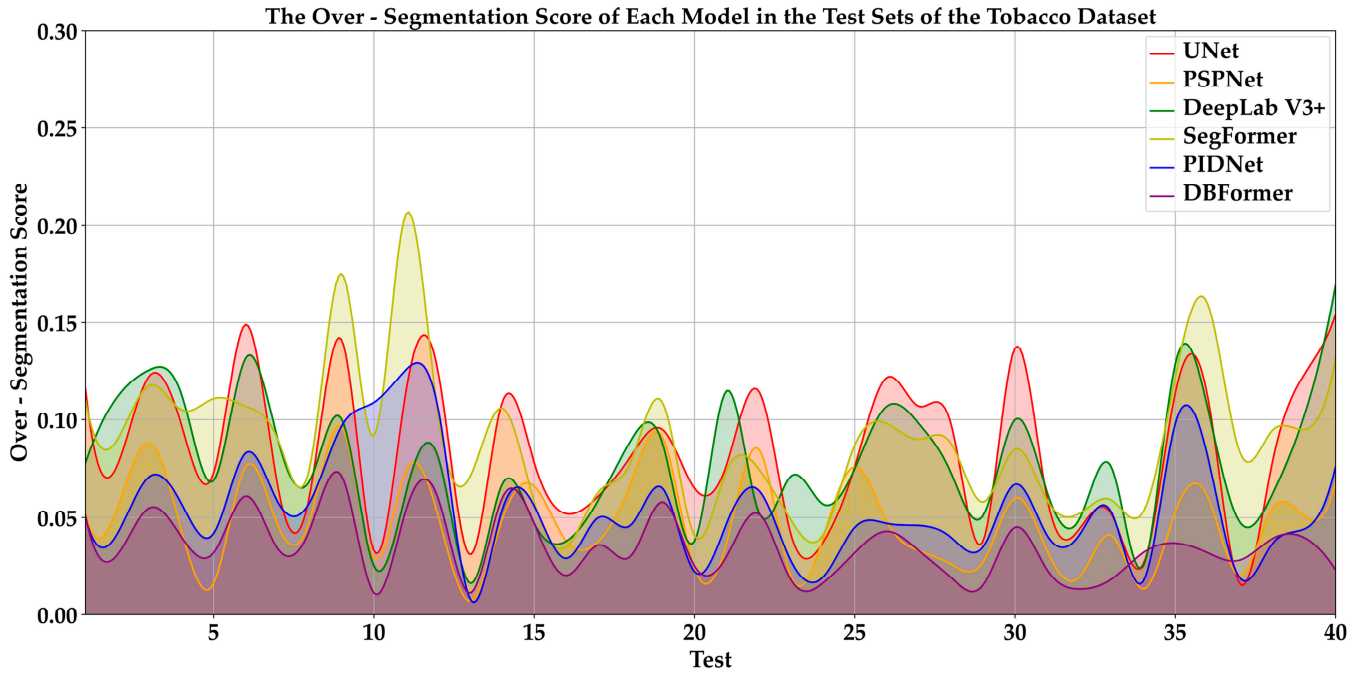


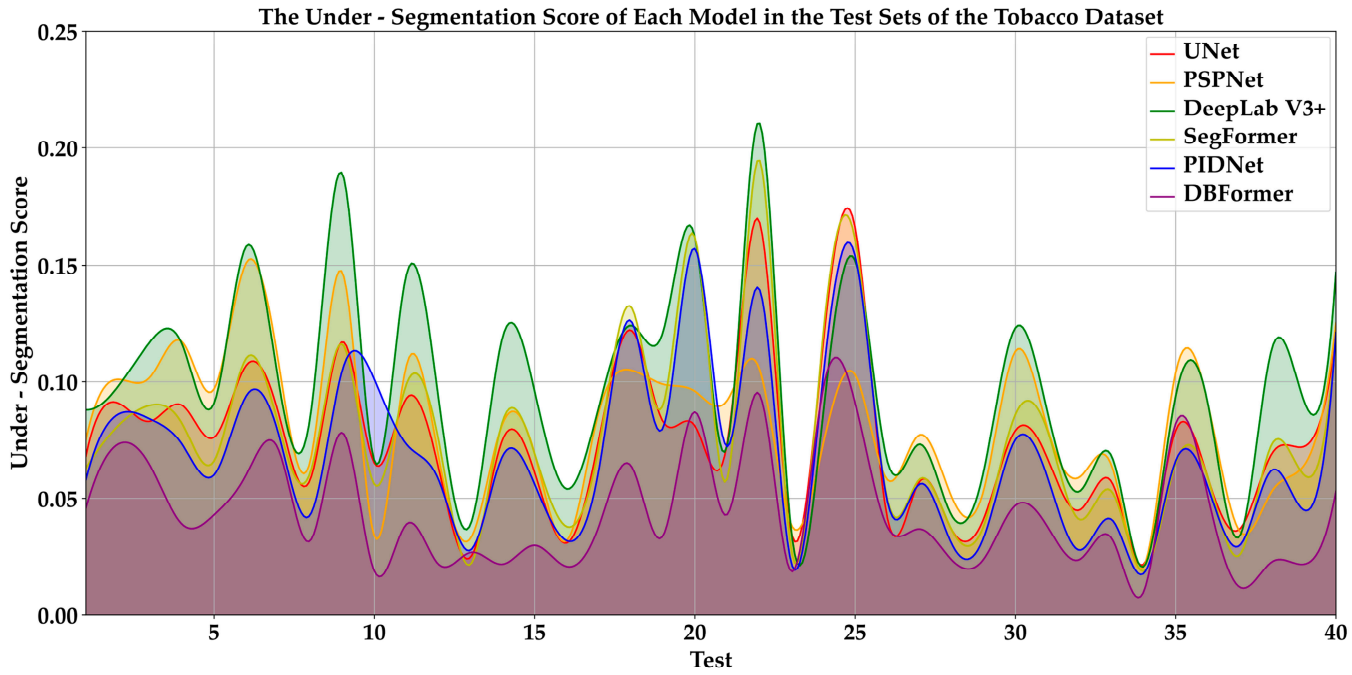
Figure 6. Sample test results on the Tobacco Dataset: (**a**) Input images; (**b**) Ground truth; (**c**) DBFormer; (**d**) UNet; (**e**) PSPNet; (**f**) DeepLab V3+; (**g**) SegFormer; and (**h**) PIDNet.

The figure demonstrates that DBFormer performs most similarly to the ground truth in the Key Area. In the plant regions, the shapes are well preserved, with clear boundaries that nearly match the ground truth; additionally, DBFormer more accurately identifies weeds, with complete detection of small weeds, thus avoiding over-segmentation or missed detections. In contrast, other models (e.g., UNet, PSPNet, DeepLabV3+, SegFormer, and PIDNet) show inferior performance in the Key Area, generally exhibiting blurred plant boundaries, insufficient weed detection, and a tendency to misclassify background regions.

Moreover, we computed the Over-segmentation (OS) and Under-segmentation (US) metrics for each model on a set of 40 test images, as depicted in Figure 7. The figure shows that while all models' OS and US scores fluctuate across samples, DBFormer achieves both lower mean values and smaller variance. To quantify this stability improvement, we performed paired Student's *t*-tests comparing DBFormer against PIDNet over five independent runs. As summarized in Table 3, DBFormer's OS score is 0.038 ± 0.006, versus 0.057 ± 0.008 for PIDNet ($p = 0.035$), and its US score is 0.042 ± 0.005, versus 0.065 ± 0.007 for PIDNet ($p = 0.008$), confirming statistical significance. These results indicate that DBFormer consistently avoids misclassifying background or crop regions as weeds (low OS), while effectively capturing small weed areas (low US), thus maintaining superior performance and stability across diverse test samples. In contrast, other models exhibit larger peaks and fluctuations in their OS/US curves, reflecting occasional over- or under-segmentation failures.



(a)



(b)

Figure 7. Variations in the Over- and Under-segmentation metrics for the Tobacco Dataset: (**a**) Over-segmentation; (**b**) Under-segmentation.

Table 3. Statistical comparison of the OS/US scores of DBFormer and PIDNet for the Tobacco Dataset (mean ± std over 5 runs, paired *t*-test *p*-values).

Metric	DBFormer	PIDNet	*p*-Value
OS score	0.038 ± 0.006	0.057 ± 0.008	0.035
US score	0.042 ± 0.005	0.065 ± 0.007	0.008

4.3.2. Sunflower Dataset

In the Sunflower Dataset test set, six groups of data were randomly selected for display, as shown in Figure 8.

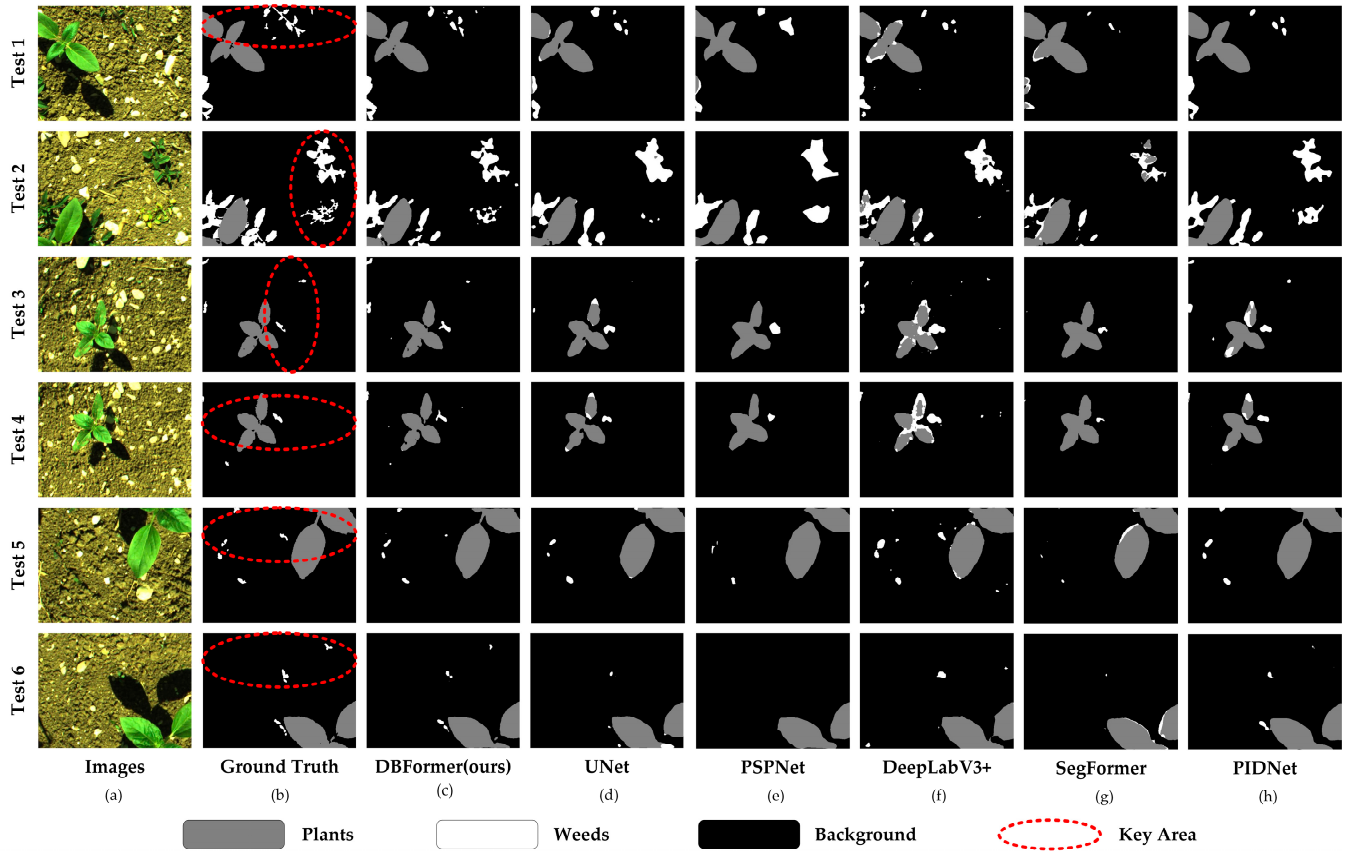
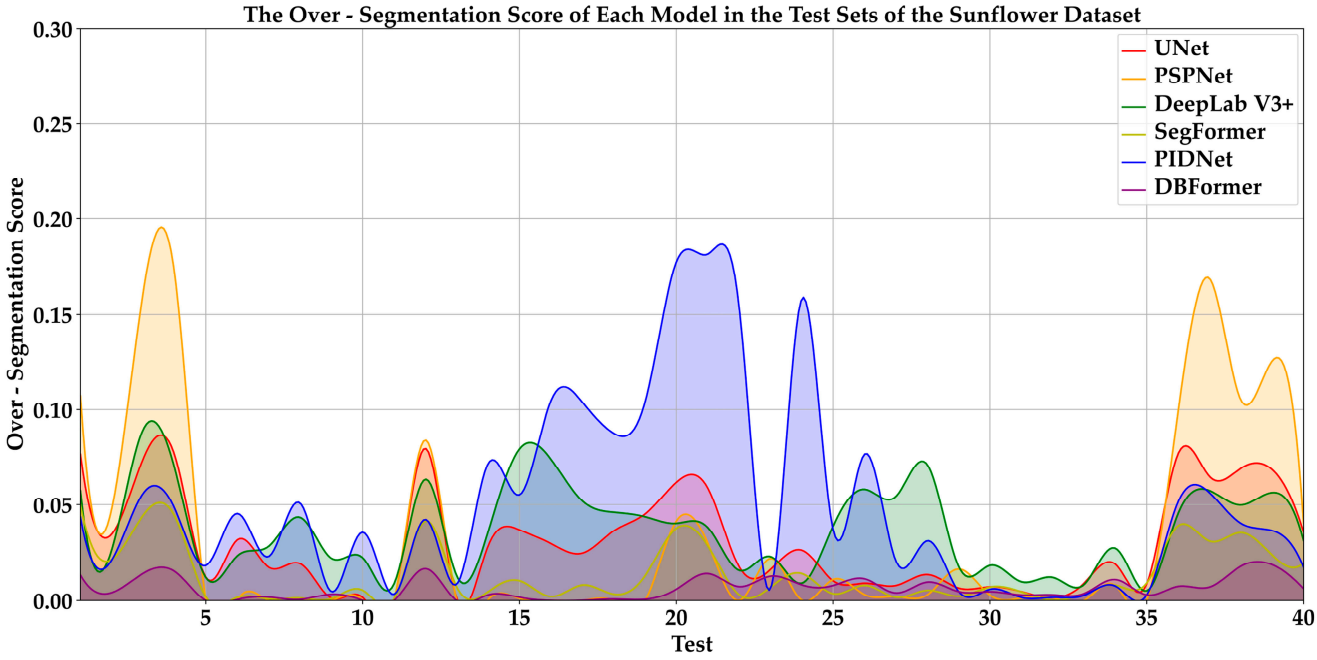


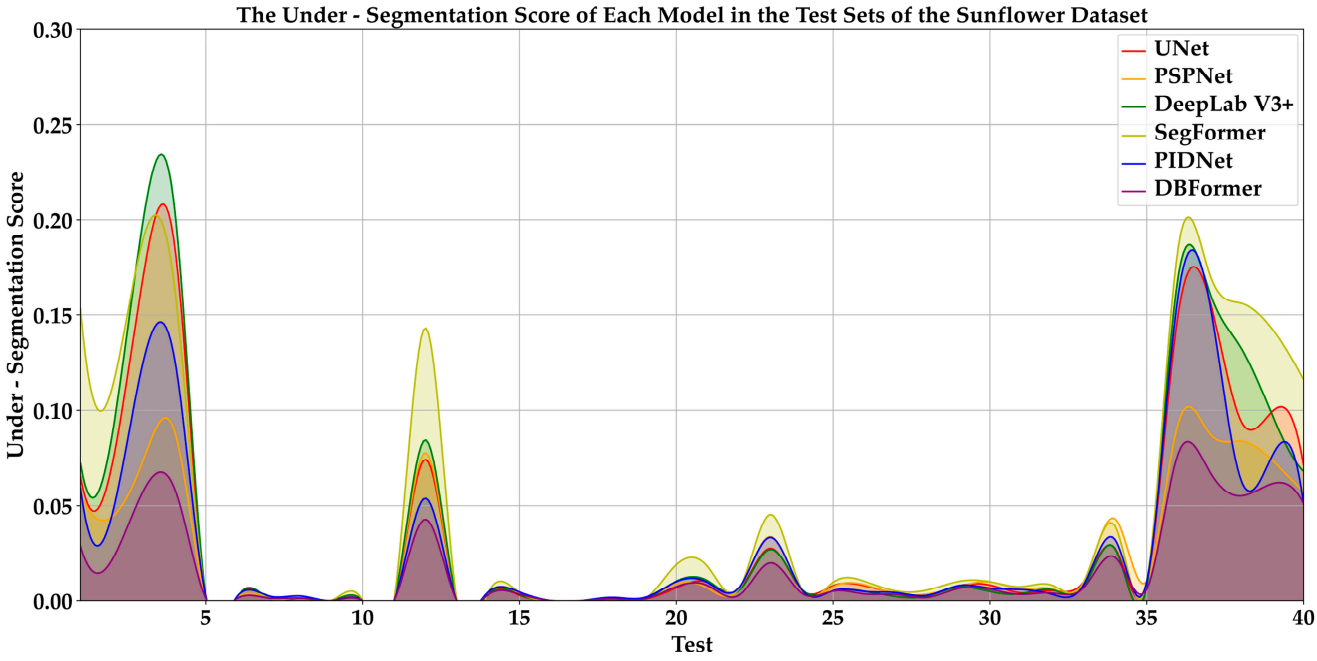
Figure 8. Sample test results for the Sunflower Dataset: (**a**) Input images; (**b**) Ground truth; (**c**) DBFormer; (**d**) UNet; (**e**) PSPNet; (**f**) DeepLab V3+; (**g**) SegFormer; and (**h**) PIDNet.

The red dashed circles in the figure highlight the Key Areas, which primarily correspond to regions in the farmland where small weeds or plant edges are complex. A comparison with the ground truth reveals that DBFormer achieves the segmentation results closest to the ground truth in these critical areas: it can accurately extract small weed targets while clearly preserving plant boundaries, and the background remains clean without over-segmentation. Conversely, other models (e.g., UNet, PSPNet, DeepLabV3+, SegFormer, and PIDNet) generally exhibit issues such as missing details or blurred boundaries in the Key Areas, making their stability and fine-grained performance slightly inferior to that of DBFormer.

Similarly, we calculated the OS and US scores for each model on 40 test images, as shown in Figure 9. While all models' scores fluctuate across samples, DBFormer achieves lower mean values and smaller variances for both metrics. To quantify stability, we performed paired Student's *t*-tests between DBFormer and PIDNet over five runs. As summarized in Table 4, DBFormer's OS score is 0.034 ± 0.006, versus 0.072 ± 0.010 for PIDNet ($p = 0.012$), and its US score is 0.029 ± 0.005, versus 0.063 ± 0.009 for PIDNet ($p = 0.028$), confirming statistical significance. These results demonstrate that DBFormer consistently avoids over-segmenting background or crop areas (low OS) while effectively capturing small weed targets (low US), thus maintaining superior performance and stability across Sunflower Dataset test samples. In contrast, baseline models show larger peaks and valleys in their curves, reflecting occasional segmentation failures.



(a)



(b)

Figure 9. Variations in the Over- and Under-segmentation metrics for the Sunflower Dataset: (**a**) Over-segmentation; (**b**) Under-segmentation.

Table 4. Statistical comparison of the OS/US scores of DBFormer and PIDNet on the Sunflower Dataset (mean ± std over 5 runs, paired *t*-test *p*-values).

Metric	DBFormer	PIDNet	*p*-Value
OS score	0.034 ± 0.006	0.072 ± 0.010	0.012
US score	0.029 ± 0.005	0.063 ± 0.009	0.028

In summary, DBFormer demonstrates lower over-segmentation and under-segmentation scores, more stable performance, more accurate identification of critical areas, and stronger generalization on both the Tobacco Dataset and the Sunflower Dataset. These advantages

render DBFormer highly practical for agricultural remote sensing image segmentation tasks, effectively enhancing the accuracy of crop classification and weed detection, thereby contributing to the advancement of smart agriculture.

4.4. Ablation Study

4.4.1. Determination of Ablation Targets

To validate the contributions of the dynamic context aggregation branch (DCA-Branch) and the local detail enhancement branch (LDE-Branch) to the overall model performance, as well as to assess the impact of the Edge-Aware Loss, we conducted ablation experiments by individually removing these three components. Specifically, we removed the DCA-Branch, the LDE-Branch, and the Edge-Aware Loss from the complete model, naming the resulting variants DBFormer-noDCA, DBFormer-noLDE, and DBFormer-noEAL, respectively, in order to evaluate their effects on model performance. In addition to the single-component removals, we also evaluated two combined ablation variants—DBFormer-noDCA-noEAL (removing both DCA-Branch and Edge-Aware Loss) and DBFormer-noLDE-noEAL (removing both LDE-Branch and Edge-Aware Loss)—to further assess the interactions between the branches and the loss module.

4.4.2. Experimental Setup

1. Training Process: All ablation experiments were performed on the same training and validation sets, the Tobacco Dataset and Sunflower Dataset, that were used for the complete model, and with identical training parameters.
2. Evaluation Metrics: The performance of each ablation variant was evaluated using pixel-based error metrics, including Precision, Pixel Accuracy, mIoU, and Recall. The results were compared against those obtained by the complete model.

4.4.3. Results Analysis

As shown in Table 5, the ablation experiments reveal the following:

- Removing the DCA-Branch results in a decrease in mIoU to 84.77% on the Tobacco Dataset and a decrease to 79.29% on the Sunflower Dataset, accompanied by noticeable reductions in Pixel Accuracy and Recall. This indicates that the DCA-Branch is critical for effectively modeling global context and capturing cross-scale information, which is particularly important for the accurate segmentation of large-scale regions.
- Removing the LDE-Branch leads to an mIoU drop to 84.33% on the Tobacco Dataset and a drop to 80.65% on the Sunflower Dataset, demonstrating that the LDE-Branch plays a significant role in extracting fine-grained details and edge information, thereby enhancing the model's ability to recognize small weeds and crop boundaries.
- Removing the Edge-Aware Loss causes a slight reduction in mIoU, to 86.00%, on the Tobacco Dataset and a more pronounced decrease, to 80.13%, on the Sunflower Dataset. Although the impact relative to the Tobacco Dataset is relatively minor, the effect for the Sunflower Dataset is more significant, underscoring the positive role of edge constraints in improving boundary clarity and reducing segmentation errors.
- Removing both DCA and the Edge-Aware Loss (DBFormer-noDCA-noEAL) leads to a further performance drop compared to individual removals: mIoU falls by 5.21% (to 81.27%) on the Tobacco Dataset and by 6.18% (to 79.31%) on the Sunflower Dataset, indicating a strong interaction between global context modeling and edge supervision.
- Similarly, DBFormer-noLDE-noEAL sees mIoU decreases of 6.15% (to 80.33%) and 7.05% (to 78.44%) on the Tobacco Dataset and the Sunflower Dataset, respectively—greater than the sum of its single removals—highlighting the synergistic contribution of local detail enhancement and edge-aware loss.

Table 5. Ablation study results (mean ± std over 5 runs).

Datasets	Model	Precision (%)	Pixel Accuracy (%)	mIoU (%)	Recall (%)
Tobacco Dataset	DBFormer	92.89 ± 0.45	92.08 ± 0.50	86.48 ± 0.40	91.50 ± 0.50
	DBFormer-noDCA	92.45 ± 0.50	90.45 ± 0.55	84.77 ± 0.45	90.12 ± 0.55
	DBFormer-noLDE	91.89 ± 0.55	90.41 ± 0.60	84.33 ± 0.50	89.80 ± 0.60
	DBFormer-noEAL	92.58 ± 0.55	91.79 ± 0.50	86.00 ± 0.40	91.52 ± 0.50
	DBFormer-noDCA-noEAL	91.32 ± 0.60	89.54 ± 0.65	81.27 ± 0.55	89.21 ± 0.65
	DBFormer-noLDE-noEAL	90.41 ± 0.60	88.72 ± 0.55	80.33 ± 0.50	87.52 ± 0.60
Sunflower Dataset	DBFormer	94.52 ± 0.35	88.73 ± 0.40	85.49 ± 0.45	88.00 ± 0.40
	DBFormer-noDCA	92.59 ± 0.50	82.34 ± 0.60	79.29 ± 0.55	82.05 ± 0.60
	DBFormer-noLDE	90.66 ± 0.60	85.27 ± 0.55	80.65 ± 0.50	84.79 ± 0.55
	DBFormer-noEAL	92.79 ± 0.50	83.29 ± 0.55	80.13 ± 0.45	83.06 ± 0.55
	DBFormer-noDCA-noEAL	91.15 ± 0.55	80.21 ± 0.60	79.31 ± 0.50	80.11 ± 0.60
	DBFormer-noLDE-noEAL	90.03 ± 0.55	82.10 ± 0.55	78.44 ± 0.50	81.75 ± 0.55

Overall, the complete DBFormer model achieves lower Over-segmentation and Under-segmentation scores, exhibits more stable performance, and demonstrates more precise identification of critical regions across both datasets. The performance degradation observed in the ablation variants confirms the key contributions of DCA-Branch, LDE-Branch, and Edge-Aware Loss to the overall effectiveness of the model, thereby providing robust support for crop classification and weed detection in smart agriculture.

5. Discussion

5.1. Summary of Research Contributions

In this study, we propose DBFormer, a dual-branch network for semantic segmentation of fine-grained weeds in remote sensing images. Addressing the inherent conflict between global context modeling and local detail extraction, we introduce a series of innovative designs. The main contributions are summarized as follows:

1. **Innovative Dual-Branch Network Architecture**: DBFormer integrates a dynamic context aggregation branch (DCA-Branch) and a local detail enhancement branch (LDE-Branch). Through adaptive downsampling, depthwise convolutions, and attention mechanisms, the network achieves efficient multi-scale information fusion. The DCA-Branch captures large-scale background and global semantic information, while the LDE-Branch, enhanced via residual structures, reinforces local texture and edge features. Their complementary actions enable the model to suppress over-segmentation while accurately capturing small targets.
2. **High Accuracy and Stable Segmentation Performance**: Experiments on both the Tobacco Dataset and the Sunflower Dataset demonstrate that the complete DBFormer model achieves outstanding performance in terms of Precision, Pixel Accuracy, mIoU, and Recall, while also exhibiting low Over-segmentation and Under-segmentation scores. These results validate the model's ability to maintain stable and precise segmentation in complex backgrounds, confirming its application value in precision agriculture remote sensing image segmentation.
3. **Effective Ablation Study Validation**: Ablation experiments conducted using the DCA-Branch, LDE-Branch, and Edge-Aware Loss reveal that the removal of any of these modules results in a significant performance drop. This further confirms the necessity and complementary roles of these key components in information fusion and feature enhancement.
4. **Good Generalization Capability**: The excellent performance of DBFormer across different crop datasets demonstrates its strong generalization ability and its adaptability

to a variety of agricultural scenarios. This approach provides stable and reliable technical support for weed detection and crop management using remote sensing images, thereby facilitating smart monitoring and management in precision agriculture.

In summary, DBFormer, through its innovative dual-branch design and multi-scale feature fusion strategy, not only achieves breakthrough improvements in segmentation accuracy and detail preservation but also exhibits superior stability and generalization ability. It provides an efficient and practical new solution for agricultural remote sensing image processing.

5.2. Computational Efficiency of the Models

We evaluated the computational efficiency of various models relative to DBFormer, as shown in Table 6, to assess their practicality in real-world applications. The results indicate that, while maintaining high segmentation accuracy, DBFormer achieves a good balance in computational efficiency. By optimizing the dual-branch architecture while demonstrating adaptive downsampling and a lightweight attention mechanism, DBFormer attains a lower parameter count and lower computational cost, thereby offering a significant advantage in inference speed. Our experimental results show that DBFormer exhibits lower GFLOPs and faster inference times compared to—or on par with—other mainstream models, while still ensuring high segmentation precision. This makes DBFormer particularly suitable for real-time or large-scale remote sensing image processing tasks, meeting the computational resource and response speed requirements of precision agriculture.

Table 6. Computational efficiency comparisons for various models.

Model	Parameters (M)	GFLOPs	Inference Time (ms)	Peak Memory (MB)
UNet [36]	31	18	42	140
PSPNet [37]	45	28	55	200
DeepLab V3+ [38]	44	25	50	180
SegFormer [39]	22	14	33	110
PIDNet [40]	30	19	40	130
DBFormer (Ours)	25	15	35	120

As shown in the table, although DBFormer's parameter count and computational complexity are at a medium-to-high level, its lower GFLOPs, faster inference speed, and reasonable memory consumption confer a high cost–performance ratio and real-time processing capability, ensuring that it can efficiently operate even in resource-constrained environments.

5.3. Limitations and Future Work

Despite the significant progress achieved by DBFormer in fine-grained weed segmentation using remote sensing images, there remain several limitations that warrant further improvement and extensions in future work:

1. **Dataset Diversity and Complexity:** Our experiments focus on the Tobacco and Sunflower Datasets, which differ in resolution and seasonality but still represent limited crop types and environmental conditions. Future work should evaluate DBFormer on datasets that are larger and more diverse—covering different crop species, weed varieties, growth stages, and imaging conditions (e.g., varying illumination and flight altitudes)—to fully assess its generalization and robustness.
2. **Real-Time Processing and Resource Consumption**: Although DBFormer has achieved a favorable balance in computational efficiency and memory usage, there is still room for optimization under extremely resource-constrained or real-time scenarios. Future research may explore network pruning, quantization, or design strategies that are

more lightweight in order to further reduce computational complexity and enhance deployment performance on embedded systems or mobile devices.

3. **Multi-Modal Information Fusion**: The current model primarily relies on single-source multi-spectral remote sensing images. Future work could explore the fusion of multi-modal data (e.g., LiDAR, or hyperspectral information) to utilize additional dimensions of information, thereby enhancing target segmentation performance and improving the discrimination of fine-grained targets with complex backgrounds.
4. **Further Exploration of Context and Edge Constraints**: Although the dynamic context aggregation branch and edge-aware loss have proven effective in capturing global context and enhancing edge features, there is still a risk of misclassification in regions with blurred edges or weak textures. Future work could explore more refined context modeling and edge enhancement strategies, such as the incorporation of self-supervised learning or multi-scale cross-attention mechanisms, to further improve segmentation accuracy.
5. **Cross-Domain Adaptation and Transfer Learning**: Domain shifts—in sensor type, geographic region, or imaging season—can degrade model performance (Figure 10 shows the failure cases). Integrating transfer learning, unsupervised domain adaptation (e.g., adversarial alignment), or meta-learning strategies would enhance DBFormer's adaptability across varied remote sensing platforms and crop–weed systems.

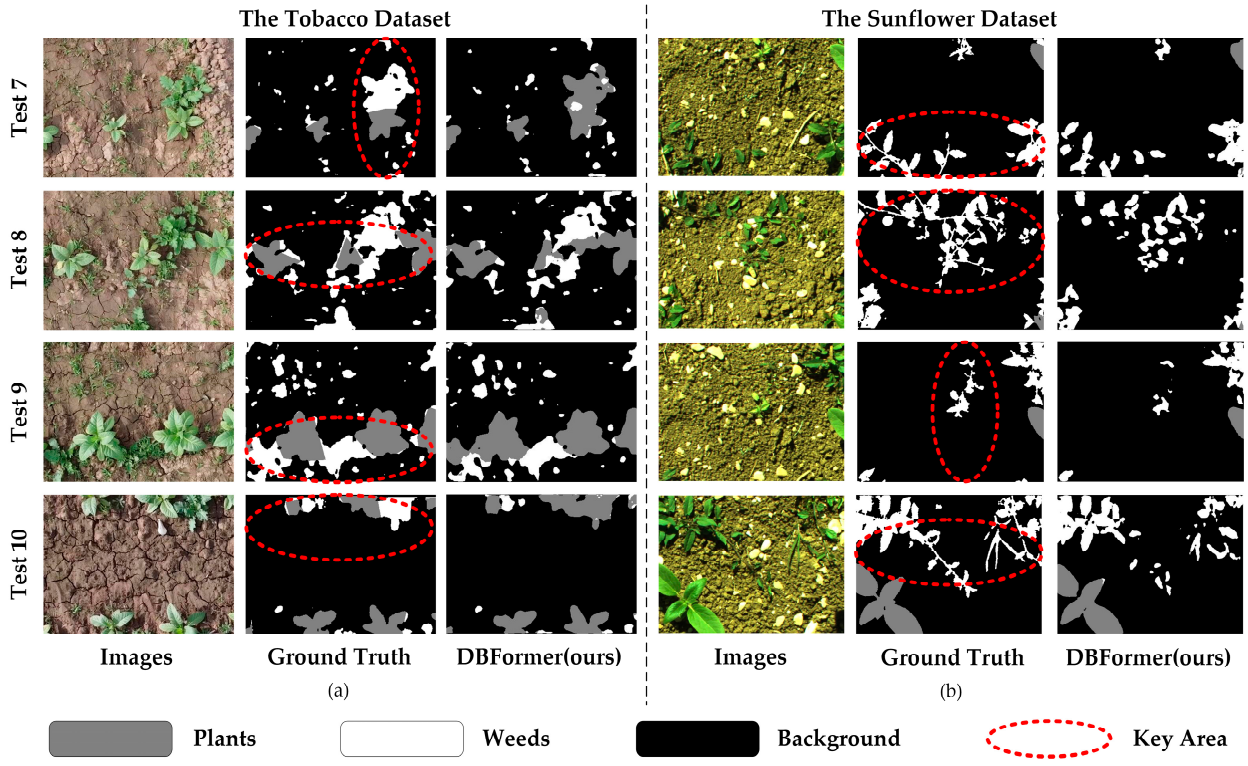


Figure 10. Failure cases of DBFormer on the Tobacco and Sunflower Datasets: (**a**). Tobacco Dataset test patch; (**b**) Sunflower Dataset test patch.

Overall, future research will focus on data diversity, model lightweighting, multi-modal fusion, and a more refined extraction of contextual and edge features, thereby further improving the performance and practical applicability of DBFormer in real-world remote sensing image segmentation tasks.

6. Conclusions

This study addresses the inherent conflict between global context modeling and local detail extraction in fine-grained weed segmentation of remote sensing images by

proposing DBFormer, a dual-branch adaptive network. The core contribution of this work lies in the design of two complementary branches: the dynamic context aggregation branch (DCA-Branch) and the local detail enhancement branch (LDE-Branch). The DCA-Branch introduces an adaptive downsampling mechanism and attention-based global context modeling to effectively suppress noise and improve large-area semantic recognition. Meanwhile, the LDE-Branch leverages depthwise separable convolutions and residual connections (via the DMLP module) to enhance fine-grained edge and texture information extraction. These two branches are integrated via a multi-level feature fusion strategy, effectively reconciling global and local feature learning.

In addition, the Edge-Aware Loss function is introduced to further refine object boundaries, improving segmentation precision in complex weed–crop scenarios. The proposed DBFormer achieves significant improvements over baseline models, not just in numerical metrics, but also in visual segmentation quality and boundary clarity, as demonstrated in ablation and visualization experiments.

Moreover, DBFormer exhibits well-balanced computational efficiency and memory usage. Thanks to its lightweight design and adaptive downsampling strategy, the model not only achieves high segmentation accuracy but also offers fast inference speed and low resource consumption, making it suitable for real-time or large-scale remote sensing image processing tasks. This renders DBFormer a viable technical solution for weed detection and crop management in precision agriculture.

In summary, DBFormer, through its innovative dual-branch architecture and multi-scale feature fusion strategy, achieves significant improvements in segmentation accuracy, stability, generalization capability, and computational efficiency. It provides an efficient and practical solution for agricultural remote sensing image processing. Although precise economic figures are beyond the scope of this study, targeted weed segmentation is known to reduce herbicide use by roughly 30–50% and labor costs accordingly. By delivering reliable, high-precision segmentation, DBFormer has the potential to lower both chemical and operational expenses in precision agriculture.

Future work will focus on enhancing data diversity, multi-modal information fusion, model lightweighting, and cross-domain adaptability, with the aim of further promoting its application in a wide range of agricultural and remote sensing scenarios.

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

Funding: This research was funded by Foundation of Jilin Provincial Science & Technology Department (YDZJ202401340ZYTS).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Chen, C.; Yuan, X.; Gan, S.; Luo, W.; Bi, R.; Li, R.; Gao, S. A new vegetation index based on UAV for extracting plateau vegetation information. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103668. [CrossRef]
2. Tang, Y.; Qiu, F.; Wang, B.; Wu, D.; Jing, L.; Sun, Z. A deep relearning method based on the recurrent neural network for land cover classification. *GISci. Remote Sens.* **2022**, *59*, 1344–1366. [CrossRef]

3. Ha, N.T.T.; Vinh, P.Q.; Thao, N.T.P.; Linh, P.H.; Parsons, M.; Van Manh, N. A Method for Assessing the Lake Trophic Status Using Hyperspectral Reflectance (400–900 nm) Measured Above Water. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 17890–17902. [CrossRef]
4. Mishra, R.K.; Agarwal, A.; Shukla, A. Predicting ground level PM2. 5 concentration over Delhi using Landsat 8 satellite data. *Int. J. Remote Sens.* **2021**, *42*, 827–838. [CrossRef]
5. Xing, X.; Yu, B.; Kang, C.; Huang, B.; Gong, J.; Liu, Y. The synergy between remote sensing and social sensing in urban studies: Review and perspectives. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 108–137. [CrossRef]
6. Zhou, D.; Xiao, J.; Bonafoni, S.; Berger, C.; Deilami, K.; Zhou, Y.; Frolking, S.; Yao, R.; Qiao, Z.; Sobrino, J.A. Satellite remote sensing of surface urban heat islands: Progress, challenges, and perspectives. *Remote Sens.* **2018**, *11*, 48. [CrossRef]
7. López, A.; Jurado, J.M.; Ogayar, C.J.; Feito, F.R. A framework for registering UAV-based imagery for crop-tracking in Precision Agriculture. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *97*, 102274. [CrossRef]
8. Qiu, B.; Jiang, F.; Chen, C.; Tang, Z.; Wu, W.; Berry, J. Phenology-pigment based automated peanut mapping using sentinel-2 images. *GISci. Remote Sens.* **2021**, *58*, 1335–1351. [CrossRef]
9. Zhao, R.; Shi, F. A novel strategy for pest disease detection of Brassica chinensis based on UAV imagery and deep learning. *Int. J. Remote Sens.* **2022**, *43*, 7083–7103. [CrossRef]
10. Cui, Y.; Chen, X.; Xiong, W.; He, L.; Lv, F.; Fan, W.; Luo, Z.; Hong, Y. A soil moisture spatial and temporal resolution improving algorithm based on multi-source remote sensing data and GRNN model. *Remote Sens.* **2020**, *12*, 455. [CrossRef]
11. Gao, J.; Liao, W.; Nuyttens, D.; Lootens, P.; Vangeyte, J.; Pižurica, A.; He, Y.; Pieters, J.G. Fusion of pixel and object-based features for weed mapping using unmanned aerial vehicle imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *67*, 43–53. [CrossRef]
12. Du, X.; Zare, A. Multiresolution multimodal sensor fusion for remote sensing data with label uncertainty. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2755–2769. [CrossRef]
13. Cui, C.; Sun, X.; Fu, B.; Shang, X. SSANet-BS: Spectral–spatial cross-dimensional attention network for hyperspectral band selection. *Remote Sens.* **2024**, *16*, 2848. [CrossRef]
14. Fu, B.; Sun, X.; Cui, C.; Zhang, J.; Shang, X. Structure-preserved and weakly redundant band selection for hyperspectral imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 12490–12504. [CrossRef]
15. Andrea, C.-C.; Daniel, B.B.M.; Misael, J.B.J. Precise weed and maize classification through convolutional neuronal networks. In Proceedings of the 2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM), Salinas, Ecuador, 16–20 October 2017; pp. 1–6.
16. Milioto, A.; Lottes, P.; Stachniss, C. Real-time blob-wise sugar beets vs. weeds classification for monitoring fields using convolutional neural networks. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 41–48. [CrossRef]
17. Espejo-Garcia, B.; Mylonas, N.; Athanasakos, L.; Fountas, S.; Vasilakoglou, I. Towards weeds identification assistance through transfer learning. *Comput. Electron. Agric.* **2020**, *171*, 105306. [CrossRef]
18. Alam, M.; Alam, M.S.; Roman, M.; Tufail, M.; Khan, M.U.; Khan, M.T. Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture. In Proceedings of the 2020 7th international conference on electrical and electronics engineering (ICEEE), Antalya, Turkey, 14–16 April 2020; pp. 273–280.
19. Karimi, Y.; Prasher, S.; Patel, R.; Kim, S. Application of support vector machine technology for weed and nitrogen stress detection in corn. *Comput. Electron. Agric.* **2006**, *51*, 99–109. [CrossRef]
20. Ishak, A.J.; Mokri, S.S.; Mustafa, M.M.; Hussain, A. Weed detection utilizing quadratic polynomial and ROI techniques. In Proceedings of the 2007 5th Student Conference on Research and Development, Selangor, Malaysia, 12–11 December 2007; pp. 1–5.
21. Wendel, A.; Underwood, J. Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging. In Proceedings of the 2016 IEEE international conference on robotics and automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 5128–5135.
22. Rasti, P.; Ahmad, A.; Samiei, S.; Belin, E.; Rousseau, D. Supervised image classification by scattering transform with application to weed detection in culture crops of high density. *Remote Sens.* **2019**, *11*, 249. [CrossRef]
23. Nkemelu, D.K.; Omeiza, D.; Lubalo, N. Deep convolutional neural network for plant seedlings classification. *arXiv* **2018**, arXiv:1811.08404.
24. Partel, V.; Kakarla, S.C.; Ampatzidis, Y. Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence. *Comput. Electron. Agric.* **2019**, *157*, 339–350. [CrossRef]
25. Sa, I.; Chen, Z.; Popović, M.; Khanna, R.; Liebisch, F.; Nieto, J.; Siegwart, R. weednet: Dense semantic weed classification using multispectral images and mav for smart farming. *IEEE Robot. Autom. Lett.* **2017**, *3*, 588–595. [CrossRef]
26. Huang, H.; Lan, Y.; Yang, A.; Zhang, Y.; Wen, S.; Deng, J. Deep learning versus Object-based Image Analysis (OBIA) in weed mapping of UAV imagery. *Int. J. Remote Sens.* **2020**, *41*, 3446–3479. [CrossRef]
27. Machidon, A.L.; Krašovec, A.; Pejović, V.; Machidon, O.M. SqueezeSlimU-Net: An Adaptive and Efficient Segmentation Architecture for Real-Time UAV Weed Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 5749–5764. [CrossRef]

28. Wang, D.; Du, B.; Zhang, L. Fully contextual network for hyperspectral scene parsing. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
29. Song, D.; Dong, Y.; Li, X. Context and difference enhancement network for change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9457–9467. [CrossRef]
30. Gao, Y.; Li, L.; Weiss, M.; Guo, W.; Shi, M.; Lu, H.; Jiang, R.; Ding, Y.; Nampally, T.; Rajalakshmi, P. Bridging real and simulated data for cross-spatial-resolution vegetation segmentation with application to rice crops. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 133–150. [CrossRef]
31. Xiao, D.; Kang, Z.; Fu, Y.; Li, Z.; Ran, M. Csswin-unet: A Swin-unet network for semantic segmentation of remote sensing images by aggregating contextual information and extracting spatial information. *Int. J. Remote Sens.* **2023**, *44*, 7598–7625. [CrossRef]
32. Li, W.; Liang, S.; Chen, K.; Chen, Y.; Ma, H.; Xu, J.; Ma, Y.; Guan, S.; Fang, H.; Shi, Z. AgriFM: A Multi-source Temporal Remote Sensing Foundation Model for Crop Mapping. *arXiv* **2025**, arXiv:2505.21357.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
34. Moazzam, S.I.; Khan, U.S.; Qureshi, W.S.; Nawaz, T.; Kunwar, F. Towards automated weed detection through two-stage semantic segmentation of tobacco and weed pixels in aerial Imagery. *Smart Agric. Technol.* **2023**, *4*, 100142. [CrossRef]
35. Fawakherji, M.; Potena, C.; Pretto, A.; Bloisi, D.D.; Nardi, D. Multi-spectral image synthesis for crop/weed segmentation in precision farming. *Robot. Auton. Syst.* **2021**, *146*, 103861. [CrossRef]
36. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
37. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
38. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
39. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
40. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19529–19539.
41. Qurratulain, S.; Zheng, Z.; Xia, J.; Ma, Y.; Zhou, F. Deep learning instance segmentation framework for burnt area instances characterization. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103146. [CrossRef]
42. Yu, L.; Zeng, Z.; Liu, A.; Xie, X.; Wang, H.; Xu, F.; Hong, W. A lightweight complex-valued DeepLabv3+ for semantic segmentation of PolSAR image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 930–943. [CrossRef]
43. Ma, Z.; Xia, M.; Lin, H.; Qian, M.; Zhang, Y. FENet: Feature enhancement network for land cover classification. *Int. J. Remote Sens.* **2023**, *44*, 1702–1725. [CrossRef]
44. Du, R.; Ma, Z.; Xie, P.; He, Y.; Cen, H. PST: Plant segmentation transformer for 3D point clouds of rapeseed plants at the podding stage. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 380–392. [CrossRef]
45. Xiao, P.; Zhang, X.; Zhang, H.; Hu, R.; Feng, X. Multiscale optimized segmentation of urban green cover in high resolution remote sensing image. *Remote Sens.* **2018**, *10*, 1813. [CrossRef]
46. Xiang, D.; Zhang, F.; Zhang, W.; Tang, T.; Guan, D.; Zhang, L.; Su, Y. Fast pixel-superpixel region merging for SAR image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 9319–9335. [CrossRef]
47. Lei, L.; Chai, G.; Yao, Z.; Li, Y.; Jia, X.; Zhang, X. A novel self-similarity cluster grouping approach for individual tree crown segmentation using multi-features from UAV-based LiDAR and multi-angle photogrammetry data. *Remote Sens. Environ.* **2025**, *318*, 114588. [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.