

Article

MSDP-SAM2-UNet: A Novel Multi-Scale and Dual-Path Model for Wheat Leaf Disease Segmentation Based on SAM2-UNet

Shuai Liu ¹, Cong Zhang ^{1,*} and Zheng Wang ²¹ School of Electrical and Electronic Engineering, Wuhan Polytechnic University, Wuhan 430048, China; liushuai_edu_cn@163.com² School of Computer Science, Wuhan University, Wuhan 430072, China; wangzwhu@whu.edu.cn

* Correspondence: hb_wh_zc@163.com

Abstract

Wheat is one of the world's essential crops, and the presence of foliar diseases significantly affects both the yield and quality of wheat. Accurate identification of wheat leaf diseases is crucial. However, traditional segmentation models face challenges such as low segmentation accuracy, limiting their effectiveness in leaf disease control. To address these issues, this study proposes MSDP-SAM2-UNet, an efficient model for wheat leaf disease segmentation. Based on the SAM2-UNet network, we achieve multi-scale feature fusion through a dual-path multi-branch architecture, enhancing the model's ability to capture global information and thereby improving segmentation performance. Additionally, we introduce an attention mechanism to strengthen residual connections, enabling the model to precisely distinguish targets from backgrounds and achieve greater robustness and higher segmentation accuracy. The experiments demonstrate MSDP-SAM2-UNet achieves outstanding performance across multiple metrics, including pixel accuracy (PA) of 94.02%, mean pixel accuracy (MPA) of 88.44%, mean intersection over union (MIoU) of 82.43%, frequency weighted intersection over union (FWIoU) of 90.73%, Dice coefficient of 81.76%, and precision of 81.63%. Compared to the SAM2-UNet, these metrics improved by 2.04%, 2.76%, 4.1%, 2.06%, 4.9%, and 3.6%, respectively. The results validate that MSDP-SAM2-UNet have tremendous segmentation performance and offer a novel perspective for wheat leaf disease segmentation.

Keywords: image segmentation; SAM2-UNet; multi-scale feature fusion; dual-path upsampling

Academic Editor: Antonio Fernández-Caballero

Received: 29 September 2025

Revised: 23 October 2025

Accepted: 3 November 2025

Published: 5 November 2025

Citation: Liu, S.; Zhang, C.; Wang, Z. MSDP-SAM2-UNet: A Novel Multi-Scale and Dual-Path Model for Wheat Leaf Disease Segmentation Based on SAM2-UNet. *Appl. Sci.* **2025**, *15*, 11778. <https://doi.org/10.3390/app152111778>

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

Crop diseases are one of the agricultural disasters in China and also a major factor constraining the sustainable development of agriculture. Leaf image segmentation for crop diseases serves as a prerequisite for disease detection, disease type identification, and disease severity assessment. Current pest identification methods encompass manual identification, traditional machine learning, and deep learning. Manual identification offers high precision but is subjective, inefficient, and costly. Traditional machine learning is efficient and cost-effective yet relies on feature extraction quality, exhibiting unstable generalisation. Deep learning enables automatic feature extraction, delivering stable and generalisable results, and shows promise in pest identification. Computer vision-based segmentation of common crop disease areas lays the foundation for automated and intelligent pest detection.

Scholars both domestically and internationally have conducted extensive research on image segmentation for crop diseases, proposing various segmentation algorithms. Syaiful et al. [1] effectively addressed the tendency of the K-means algorithm to become trapped in local optima through a particle swarm optimisation approach, demonstrating excellent segmentation performance for tomato early blight. Beyond employing single segmentation algorithms, combining different methodologies represents a current research focus. Addressing the uneven colour distribution in rice leaf blast disease areas, Bakar et al. [2] employed a multi-threshold segmentation algorithm and edge detection within the H channel of the HSV colour space to delineate complete lesions. This approach further enabled the classification of disease severity. Jun et al. [3] applied an adaptive segmentation algorithm combining local thresholds with seed region growth for maize leaf lesion segmentation, yielding superior results compared to single algorithms. The aforementioned studies employ traditional image segmentation methods. While these approaches are straightforward to comprehend and computationally simple, they exhibit significant limitations, such as poor generalisability.

In recent years, convolutional neural networks have played a pivotal role in advancing medical image analysis [4]. In 2015, the introduction of the fully convolutional network (FCN) paved the way for deep learning applications in image segmentation [5]. That same year, Ronneberger et al. [6] introduced U-Net, a semantic segmentation framework that extends FCN with an encoder–decoder architecture. This network integrates semantic information with spatial positional information lost during downsampling through a unique skip connection structure between encoder and decoder, delivering excellent segmentation performance even for small objects. U-Net, a semantic segmentation model characterized by its simplicity and broad applicability, has achieved outstanding results in image segmentation.

In recent years, numerous scholars have applied U-Net to crop disease segmentation with remarkable results. As research deepens, U-Net models and their variants have gained widespread adoption across diverse image segmentation tasks due to their flexible network architecture and outstanding performance. For instance, Xu et al. [7] introduced a lightweight multi-scale U-Net architecture enhanced with residual connections between encoding and decoding layers. This design facilitated the integration of shallow and deep features, leading to effective segmentation of crop disease leaf images.

In April 2023, the Segment Anything Model (SAM) [8] emerged as a fundamental large-scale solution for segmentation tasks in natural images. Based on the Vision Transformer (ViT) [9] architecture, SAM integrates three unique prompting modes—point prompts, bounding box prompts, and text prompts—to achieve outstanding performance in natural image segmentation tasks [10]. Consequently, since its introduction, numerous researchers have applied the SAM to fields such as laser defect detection, remote sensing image segmentation, and medical image segmentation [11].

Li et al. [12] creatively integrated SAM's robust feature extraction with nnUNet's adaptability to propose the nnSAM, thereby improving segmentation accuracy and reliability on small datasets. Ravi et al. [13] presented an innovative framework for visual segmentation, termed Segment Anything Model 2 (SAM 2), tailored for diverse segmentation tasks in visual media such as images and videos. Li et al. [14] developed a dental segmentation approach by utilizing the pre-trained SAM 2 model. Their method incorporated an adapter module for fine-tuning, ScConv modules for multi-scale feature extraction, and gated attention mechanisms to strengthen semantic segmentation in medical imaging.

In this paper, we propose an MSDP-SAM2-UNet image segmentation algorithm for wheat leaf blight, optimizing the SAM2-UNet model's multi-branch design and performing multi-scale feature fusion to enhance the network's feature extraction capability. We

introduce an attention mechanism in skip connections and reinforce residual connections to improve model training stability. Within the decoder, dual-path Content-Aware ReAssembly of Features (CARAFE) upsampling is implemented to elevate segmentation accuracy and quality, effectively addressing diverse challenges inherent in wheat leaf blight image segmentation tasks.

Our contributions in this study are as follows:

1. Addressing the complex segmentation challenges of wheat pest and disease lesions—which typically involve details of varying sizes and scales—our ResPathWithMultiScale (RPMS) Model. By optimizing the multi-branch structure, introducing a channel attention mechanism, and enhancing residual connections, it effectively addresses the shortcomings of the RFB module in segmenting complex crop lesion areas, improving feature extraction capabilities for intricate lesions.
2. Addressing the complexity of leaf-edge lesions in wheat pests and diseases, existing decoders based on the classic U-Net architecture struggle to capture information from small, localized lesions. We introduce the DualPathModule, which employs dual-path Content-Aware Reassembly of Features (CARAFE) upsampling at high and low resolutions. This enables the model to perceive lesions across different scales effectively, thereby improving segmentation performance for crop lesions.
3. We attempted to simultaneously incorporate the ResPathWithMultiScale (RPMS) Model and DualPathModule to enhance the SAM2-UNet model, proposing the MSDP-SAM2-UNet model. Experiments revealed its effective handling of various challenges in wheat lesion segmentation tasks. By comprehensively integrating information from lesions of different scales from multiple perspectives, it accurately distinguishes targets from backgrounds, thereby improving segmentation accuracy and quality.

2. Methods

2.1. SAM2-UNet

SAM2 (Segment Anything 2) [13] is a novel model introduced by Meta, designed to segment any content within an image without being restricted to specific categories or domains. U-Net, introduced by Ronneberger et al. [6] in 2015, is a U-shaped convolutional architecture featuring a symmetric encoder–decoder framework. SAM2-UNet is a model integrating SAM2 and U-Net, leveraging SAM2’s feature extraction capabilities alongside U-Net’s upsampling and feature fusion mechanisms to achieve efficient image segmentation.

SAM2-UNet employs the pre-trained Hiera backbone from SAM2, which captures hierarchical multi-scale features, making it well-suited for designing U-shaped networks [15]. The Receptive Field Block (RFB) feature fusion module further processes extracted features. RFB enhances the network’s feature extraction capabilities by simulating the receptive fields of human vision.

Each adapter in SAM2-UNet comprises a linear layer for downsampling and upsampling, with GeLU activation functions in between. Unlike SAM2’s bidirectional transformer decoder, the SAM2-UNet decoder adopts a classic U-Net design, consisting of three decoder modules. Each module contains two pairs of “Conv-BN-ReLU” components, featuring a 3×3 convolutional layer, batch normalization, and ReLU activation. The decoder of U-Net is used to gradually restore the image resolution through up-sampling and feature concatenation, thereby generating segmentation masks. The up-sampling module achieves feature up-sampling through bilinear interpolation and feature fusion. Side outputs are added at different feature levels to generate multi-scale segmentation results. Each segmentation output is compared with the ground-truth segmentation mask to calculate the loss, and deep supervision is applied.

2.2. MSDP-SAM2-UNet

The segmentation results generated by the SAM2-UNet model are class-agnostic. Without manual prompts for specific categories, SAM2 cannot produce segmentation results for designated classes. To enhance the specificity and better adapt it to specific downstream tasks while efficiently utilizing pre-trained parameters, we propose the MSDP-SAM2-UNet architecture (as shown in Figure 1).

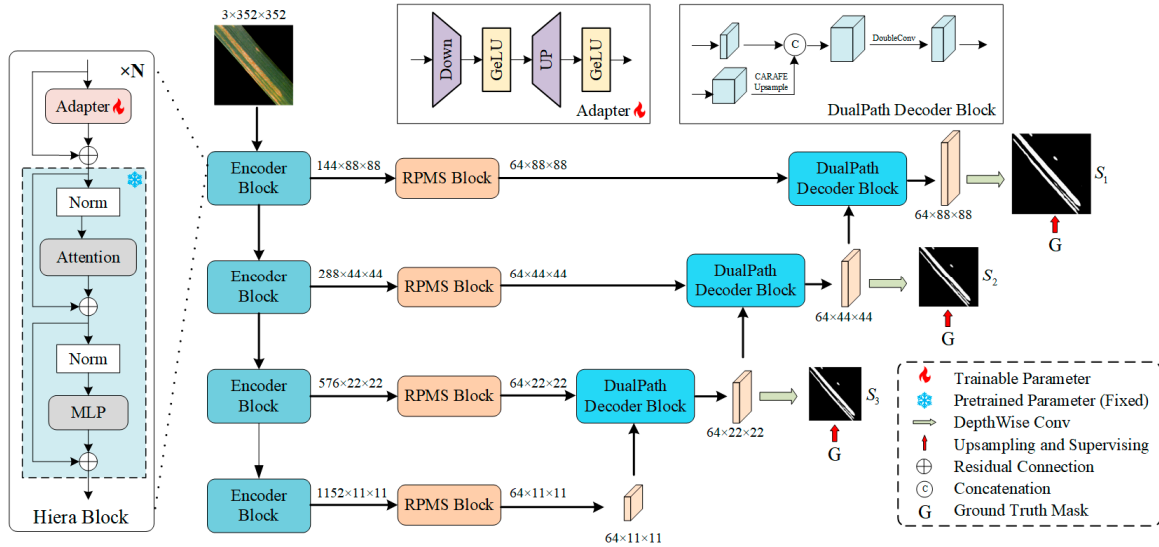


Figure 1. Overview of the proposed MSDP-SAM2-UNet.

Encoder: MSDP-SAM2-UNet adopts the SAM2-pretrained Hiera-L backbone. Its attention mechanisms overcome CNNs' long-range contextual limitations, and the Hiera Block's hierarchy aids multi-scale feature extraction, fitting U-shaped networks. For parameter-efficient fine-tuning, a trainable adapter module is introduced before the Hiera Block, with the Hiera Block's parameters kept frozen. This approach eliminates the need to fine-tuning the Hiera Block, significantly reducing memory usage.

Given an input image $I \in \mathbb{R}^{3 \times H \times W}$, where H and W denote the image's height and width, the Hierarchical (Hier) backbone in our model generates four levels of hierarchical features. These features are represented as $X_i \in \mathbb{R}^{C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ ($i \in \{1, 2, 3, 4\}$) with the channel counts at each level being ($C_i \in \{144, 288, 576, 1152\}$).

Adapter: The adapter module, inspired by Houlsby et al. [16] and Qiu et al. [17], adopts a sequential framework composed of a downsampling linear layer, followed by a GeLU activation function, an upsampling linear layer, and a subsequent GeLU activation function. This configuration enables efficient fine-tuning of the Hiera Block while minimizing memory usage.

The RPMS Block: After completing feature extraction at the encoder stage, the channel count is reduced to 64 utilizing receptive field blocks [18]. Subsequently, multi-channel feature integration is executed via deep convolution operations. The module's multi-branch architecture employs convolutional kernels of varying sizes to simulate the multi-scale receptive field characteristics of human vision. Combined with residual connections, this approach extracts feature information across different scales.

The DualPath Decoder Block: We modified the traditional U-Net decoder by retaining the up-sampling operations. Drawing on the U-Net's highly adaptable structure [19], we implemented a customized DoubleConv module composed of two identical "Conv-BN-ReLU" combinations, with the convolution operation utilizing a 3×3 kernel. Each decoder's output features are processed through a 1×1 convolutional segmentation head

to generate segmentation results $S_i \in \{1, 2, 3\}$. These results are then upsampled and supervised based on ground-truth segmentation masks.

Loss Function: Each hierarchical loss function in MSDP-SAM2-UNet consists of a weighted union of the Intersection over Union (IoU) and Binary Cross-Entropy (BCE) losses [20]. The specific single-level loss function is defined as follows:

$$L = L_{IoU}^{\omega} + L_{BCE}^{\omega} \quad (1)$$

Since we employ deep supervision, the final loss function of MSDP-SAM2-UNet is expressed as the sum of individual hierarchical losses:

$$L_{total} = \sum_{i=1}^3 L(G, S_i) \quad (2)$$

2.2.1. ResPathWithMultiScale Block

In this study, to address the varying sizes and shapes of crop lesion areas, we replaced the RFB in SAM2-UNet with the ResPathWithMultiScale Block (RPMS Block). The RPMS Block extends and optimizes the multi-branch design of the RFB, as shown in Figure 2. This flowchart illustrates a deep learning-based convolutional neural network (CNN) architecture, with the Receptive Field Block (RFB) at its core. The RFB module employs a multi-branch design combining convolutional kernels of varying sizes and dilated convolutions to enhance the network's receptive field and capture multi-scale features. Input data first enters the RFB module, where it undergoes processing through multiple branches. Each branch employs convolutional operations with distinct parameters (e.g., varying kernel sizes, different dilation rates) to capture features across scales and orientations. Outputs from these branches are concatenated and fused, then processed through 1×1 convolutions and activation functions for further feature integration. Finally, the Channel SE module enhances the expression of key features, producing the processed feature map.

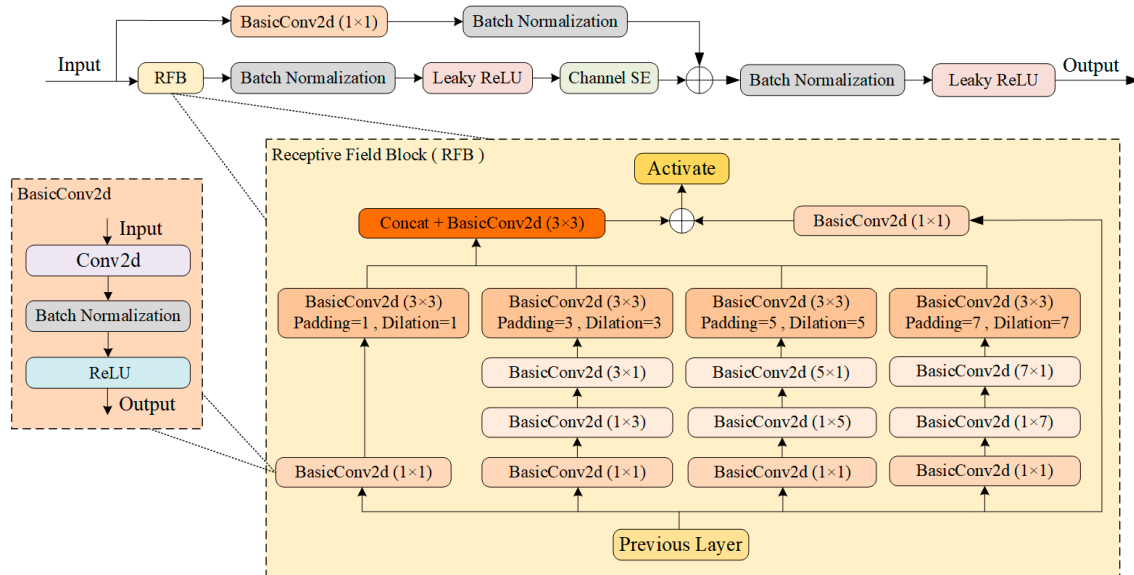


Figure 2. The structure of RPMS Block. This module primarily leverages the multi-branch design of the RFB block to enhance the network's receptive field and achieve multi-scale feature extraction.

In the RFB, multiple branches with different dilation rates expand the receptive field and capture global contextual information. The 3×3 convolution layer extracts local features. There are four branches with different padding and dilation rates:

- Branch 0: Uses standard convolution to extract local boundary features, such as fine details of disease areas.
- Branch 1: Employs dilated convolution (dilation = 3) to capture larger contextual information, aiding in identifying the shape of disease damages.
- Branch 2: Utilizes a larger dilation rate (dilation = 5) to further expand the receptive field for extracting broader contextual information.
- Branch 3: Implements the largest dilation rate (dilation = 7) to capture global contextual information, helping to identify the distribution of disease areas in the image.

Through multi-branch and dilated convolutions within the RFB module, multi-scale feature extraction is performed. Dilated convolutions expand the receptive field, making it suitable for tasks requiring global context (such as object detection and semantic segmentation). The ChannelSE attention mechanism module enhances important feature channels while suppressing unimportant ones. After concatenating features from all branches, the convolutional layer fuses multi-scale features. Batch Normalization is applied multiple times to the convolutional outputs for accelerated training. This approach removes redundancy while extracting scale-related feature correlations. Additionally, residual connections preserve the original input features, aiding deep network training and mitigating the vanishing gradient problem. Finally, the Leaky ReLU activation function introduces nonlinearity to address ReLU's "neuron death" issue. The proposed design enables the model to effectively extract features at diverse scales, from minute spots to extensive decay regions, leading to enhanced segmentation accuracy of lesions. Additionally, this architecture is tailored to capture multi-scale features, rendering it especially suitable for tasks involving leaf lesions of varying sizes.

2.2.2. DualPath Decoder Block

As shown in Figure 3, to maintain high feature expression capability and computational efficiency, we designed the DualPath Decoder Block Model to replace the original module's decoder block.

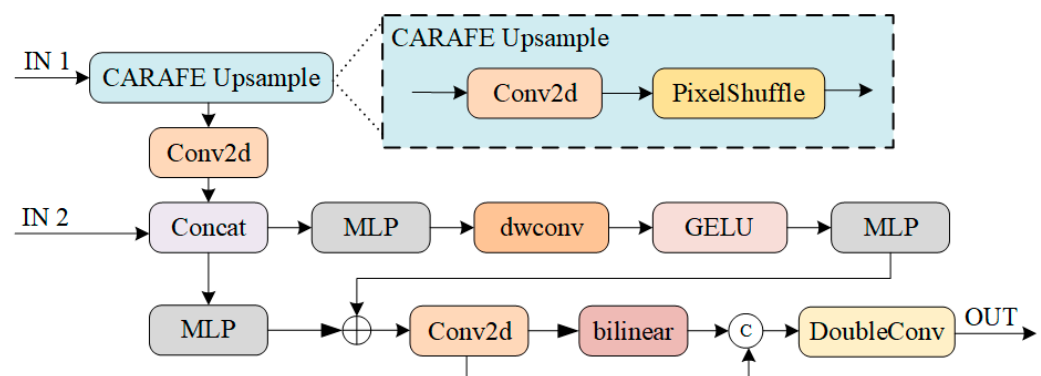


Figure 3. The structure of DualPath Decoder Block. The block utilizes CARAFE for upsampling to aggregate more contextual information, achieving feature fusion through a dual-path structure.

The DualPath Decoder Block primarily consists of: Content-Aware Reassembly of Features (CARAFE) [21] Upsample, Multilayer Perceptron (MLP), and bilinear interpolation components. In the CARAFE Upsample stage, features are first upscaled via CARAFE to generate higher-resolution feature maps. Features are then extracted through convolutional operations, followed by further upscaling via PixelShuffle to enhance resolution while concurrently extracting features. Within the MLP module, feature maps from the CARAFE Upsample module undergo concatenation (Concat) to fuse both feature sets. Subsequently, the sequence includes MLP, dwconv, GELU, and MLP stages to extract features and introduce nonlinearity, thereby enhancing feature representation. In the bilinear interpolation

stage, the output from the MLP module undergoes bilinear interpolation to adjust the feature map resolution to an appropriate size. This is then multiplied point-by-point with the output from the CARAFE Upsample module to achieve feature fusion. Finally, the Conv2d and DoubleConv layers produce the final image segmentation result.

The Multilayer Perceptron (MLP) is a neural network module composed of multiple fully connected layers. Its primary functions within the module are twofold: first, to perform nonlinear transformations on concatenated features, extracting more effective feature representations; second, to adjust the dimensionality of bilinearly interpolated features, matching them with another branch of features to facilitate subsequent dot product operations.

As shown in Figure 4, the PixelShuffle operation [22] is typically used to transform a high-dimensional feature map into a low-dimensional feature map with higher spatial resolution. The feature map is expanded in the spatial dimension by increasing its width and height, thereby achieving an upscaling effect. This process preserves the integrity of the feature information, ensuring that the feature map does not lose excessive detail during upscaling.

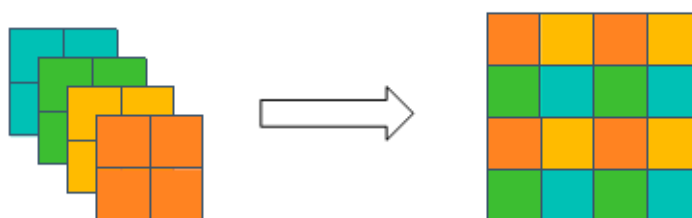


Figure 4. The example of Pixel Shuffle.

Specifically, bilinear interpolation estimates the value of a target pixel by performing a weighted average of the values of four adjacent pixels in the image. For a target pixel point, its value is calculated through interpolation using the values of the four adjacent pixels and their distances to the target point. The formula is as follows:

$$I(x, y) = \frac{(x_2 - x)}{(x_2 - x_1)} \times \frac{(y_2 - y)}{(y_2 - y_1)} \times I(x_1, y_1) + \frac{(x - x_1)}{(x_2 - x_1)} \times \frac{(y_2 - y)}{(y_2 - y_1)} \times I(x_2, y_1) \\ + \frac{(x_2 - x)}{(x_2 - x_1)} \times \frac{(y - y_1)}{(y_2 - y_1)} \times I(x_1, y_2) + \frac{(x - x_1)}{(x_2 - x_1)} \times \frac{(y - y_1)}{(y_2 - y_1)} \times I(x_2, y_2) \quad (3)$$

where $I(x, y)$ is the target pixel value, and $I(x_1, y_1), I(x_2, y_1), I(x_1, y_2), I(x_2, y_2)$ are the values of adjacent pixels.

In summary, the DualPath Decoder Block architecture leverages a dual-path structure and feature fusion extraction to delve deeper into the relationships between features, thereby enhancing their expressive power. Simultaneously, the introduction of the GELU activation function effectively avoids both vanishing gradients and exploding gradients, contributing to improved model training speed and convergence performance. This enables the model to better fit the data.

3. Experiments and Analysis

To ensure the reliability of the experimental conclusions, all experiments were conducted under identical experimental conditions. The experiments employed an NVIDIA A40 GPU with 48 GB of memory (designed by NVIDIA Corporation, located in Santa Clara, CA, USA), running the PyTorch 1.13.0 deep learning framework on the Ubuntu 22.04 operating system. The programming language used throughout the study was Python 3.9, with CUDA version 11.7. The model was built on the PyCharm 2022.3 platform.

3.1. Dataset

The experimental materials selected for this study originate from the PlantSeg dataset (<https://doi.org/10.5281/zenodo.13293891>), a large-scale segmentation dataset focused on plant diseases. The PlantSeg dataset consists of 11,400 image samples associated with 115 plant diseases across 34 plant hosts, each annotated with disease segmentation masks. Additionally, it includes 8000 healthy plant images categorized by plant type [23]. For the present study, 9163 disease images were selected as the training set, while 2295 images were allocated to the test set. These subsets were employed in the pre-training phase to address the challenge of limited training samples.

In this experiment, a pre-trained model was utilized to focus on wheat leaf disease detection, particularly targeting wheat leaf blight and leaf rust. The experimental dataset comprises 1434 training images and 358 test images, maintaining an 8:2 training-to-testing ratio. Selected wheat disease images are illustrated in Figure 5.



Figure 5. Some images of wheat leaf diseases: (a) A single leaf blight lesion on a wheat leaf (b) Multiple leaf blight lesions on a single leaf (c) Some narrow and short wheat streak rust lesions (d) Some broad and long wheat streak rust lesions.

3.2. Evaluation Metrics

In image segmentation, various widely recognized criteria are utilized to assess algorithmic performance. Below, we will introduce several commonly used pixel-level accuracy metrics, which are all based on the calculation of the Confusion Matrix.

For illustrative purposes, as shown in Table 1, assume that there are $k + 1$ classes (including one background class); L_0 to L_k . p_{ii} represents the number of pixels that belong to class i and are predicted as class i , that is, the number of correctly classified pixels for class i . p_{ij} represents the number of pixels that belong to class i but are predicted as class j , while p_{ji} represents the number of pixels that belong to class j but are predicted as class i [24]. The values of k represent the number of classes, and p_{ij} and p_{ji} are interpreted as false positives and false negatives, respectively. The calculation formula for the semantic segmentation metric used in this study is as follows.

Table 1. The classification results of confusion matrix.

Actual Situation	Predicted Result	
	Positive	Negative
Positive	True Positive (p_{ii})	False Negative (p_{ji})
Negative	False Positive (p_{ij})	True Negative (p_{jj})

Accuracy corresponds to pixel accuracy (PA), which describes the proportion of correctly predicted values in the predicted results [25], that is, the proportion of correctly

predicted pixel points in the test target relative to all pixels in the test area. The formula is Equation (4).

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (4)$$

Mean Pixel Accuracy (*MPA*) calculates the proportion of correctly classified pixels for each class [25], then averages all classes, that is, calculates the PA value for each test target, and then averages all PA values. The formula is Equation (5).

$$MPA = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}} \quad (5)$$

Mean Intersection over Union (*MIoU*) [25] is a measure of semantic segmentation, which calculates the intersection over union ratio between the predicted area and the actual area of the test target as the standard of semantic segmentation. The formula is Equation (6).

$$MIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (6)$$

Frequency Weighted Intersection over Union (*FWIoU*) [25] is based on the frequency of each class occurrence, weighted by each class's IoU and then summed. The formula is Equation (7).

$$FWIoU = \frac{1}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (7)$$

The evaluation metric of the model is *Dice* (also known as F1-Score), which is calculated based on the precision and recall of the test sample. The formula is Equation (8).

$$Dice = \frac{2 \sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji}} \quad (8)$$

Precision refers to the proportion of samples correctly predicted as positive out of all samples predicted as positive. In the context of a confusion matrix, the formula is Equation (9).

$$Precision = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (9)$$

Here, p_{ii} represents the number of pixels correctly classified, and $\sum_{i=0}^k \sum_{j=0}^k p_{ij}$ represents the total number of pixels predicted as positive.

3.3. Comparative Experiments

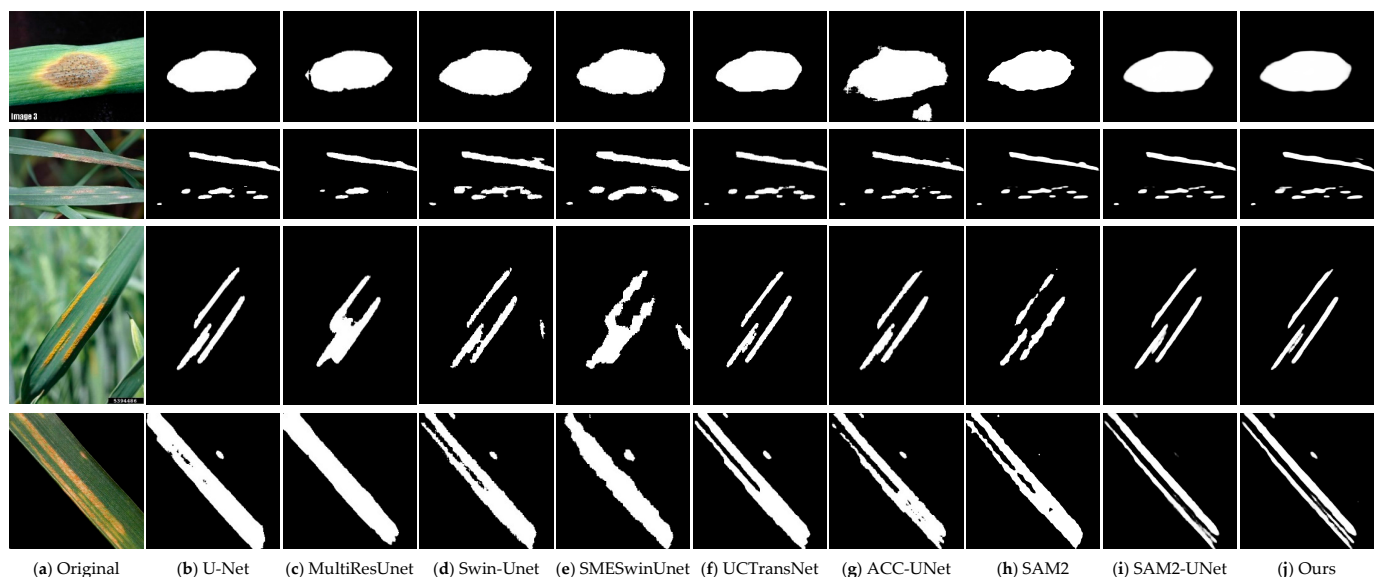
To demonstrate the efficacy of the MSDP-SAM2-UNet model, we conducted comparative analysis against eight other prominent segmentation models, including U-Net, MultiResUnet [26], Swin-Unet [27], SMESwin-Unet [28], UTransNet [29], ACC-UNet [30], SAM2, SAM2-UNet. All models were subjected to identical training and testing, and their performance was assessed by six metrics, including PA, MPA, MIoU, FWIoU, Dice, and Precision. The detailed comparative outcomes are meticulously tabulated in Table 2.

As shown in Table 2, the results of the experiments demonstrate MSDP-SAM2-UNet achieves outstanding performance across multiple metrics, including PA of 94.02%, MPA of 88.44%, MIoU of 82.43%, FWIoU of 90.73%, Dice of 81.76%, and Precision of 81.63%. Compared to the SAM2-UNet, these metrics improved by 2.04%, 2.76%, 4.1%, 2.06%, 4.9%, and 3.6%, respectively.

Table 2. Comparative experimental results.

Model	PA (%)	MPA (%)	MIoU (%)	FWIoU (%)	Dice (%)	Precision (%)
U-Net	87.48	78.48	68.90	81.65	66.24	63.05
MultiResUnet	85.50	77.61	64.64	78.84	60.71	61.79
SwinUnet	88.34	79.49	70.39	82.73	68.77	63.23
SMESwin-Unet	88.31	78.39	69.15	82.41	66.95	61.25
UCTransNet	86.77	79.55	69.07	81.05	67.37	64.14
ACC-UNet	87.32	80.14	69.32	81.47	67.64	65.08
SAM2	90.72	79.57	69.74	84.88	64.98	66.11
SAM2-UNet	91.98	85.68	78.33	88.67	76.86	78.03
Ours	94.02	88.44	82.43	90.73	81.76	81.63

To visually contrast the segmentation performance of nine models in wheat leaf disease segmentation, the detailed experimental results are presented in Figure 6. The figures reveal that MSDP-SAM2-UNet achieves superior segmentation accuracy and notable edge preservation, with results closely matching ground truth. In contrast, other models are prone to over-segmentation, under-segmentation, or edge blurring.

**Figure 6.** The segmentation results of different methods.

Notably, the SMESwinUnet and MultiResUnet models show particularly significant classification errors for streak rust lesions. These results demonstrate that the proposed MSDP-SAM2-UNet model maintains outstanding segmentation performance even when confronted with complex leaf lesion images.

3.4. Ablation Experiments

To evaluate the effectiveness of each component in the model, we designed four sets of ablation experiments, with results shown in Table 3. Under identical experimental settings, we first replaced the RFB module with the ResPathWithMultiScale Block (RPMS Block). Compared to the baseline model, the improved model achieved increases of 0.7%, 1.1%, 1.24%, 0.28%, 2.1%, and 0.71% in PA, MPA, MIoU, FWIoU, Dice, and Precision, respectively. This demonstrates that the proposed RPMS module enhances the model's segmentation performance by expanding the receptive field through a multi-branch architecture and convolutions with varying dilation rates, thereby capturing global contextual information and strengthening the model's ability to extract local image features.

Table 3. The results of ablation experiments. Using SAM2-UNet as the baseline model, the results demonstrate that the RPMS and DualPath blocks deliver performance improvements.

Model	PA (%)	MPA (%)	MIoU (%)	FWIoU (%)	Dice (%)	Precision (%)
Baseline	91.98	85.68	78.33	88.67	76.86	78.03
Baseline + RPMS	92.68	86.78	79.57	88.95	78.96	78.74
Baseline + DualPath	93.23	87.21	81.49	89.28	79.97	79.86
Baseline + RPMS + DualPath (Ours)	94.02	88.44	82.43	90.73	81.76	81.63

Subsequently, we introduced a novel DualPathModule decoder module. Compared to the baseline model, the improved model achieved increases of 1.25%, 1.53%, 3.16%, 0.61%, 3.11%, and 1.83% in PA, MPA, MIoU, FWIoU, Dice, and Precision, respectively. This demonstrates that the proposed DualPathModule decoder, by utilizing the CARAFE module for upsampling combined with bilinear interpolation, processes input features through two distinct paths. This approach fuses multi-scale information to more accurately recover subtle edge details in images, significantly enhancing segmentation precision and accuracy.

Finally, we combined the ResPathWithMultiScale Block (RPMS Block) module with the DualPathModule decoder module for experiments under identical conditions. Compared to the baseline model, the improved model achieved increases of 2.04%, 2.76%, 4.1%, 2.06%, 4.9%, and 3.6% in PA, MPA, MIoU, FWIoU, Dice, and Precision, respectively. Experimental data reveal that improvements in PA, MPA, and MIoU are nearly equivalent to the combined effects of the ResPathWithMultiScale Block (PMRS Block) and DualPathModule decoder acting independently. Notably, FWIoU, Dice, and Precision achieve significantly better results than either component alone. Collectively, these findings demonstrate that our proposed model delivers superior segmentation performance compared to the baseline model.

4. Discussion

This study focuses on wheat leaf blight and streak rust. The proposed MSDP-SAM2-UNet model demonstrates outstanding performance in wheat leaf disease segmentation tasks, addressing the limitations of existing models in segmenting multiple lesions and detecting minute elongated lesions on wheat leaves. Ablation experiments demonstrated that the RPMS Block and DualPath Decoder Block significantly enhanced the model's segmentation performance. Compared to the baseline model, the proposed improvements effectively enhanced the segmentation of wheat leaf lesions, resulting in a significant overall performance boost.

Furthermore, the comparative experimental outcomes highlighted the MSDP-SAM2-UNet model's exceptional superiority. Comparative analysis with other classical segmentation models led to the following conclusions: (1) Compared with the SAM2-UNet baseline model, the PA, MPA, MIoU, FWIoU, Dice and Precision of the improved model increased by 2.04%, 2.76%, 4.1%, 2.06%, 4.9% and 3.6%, respectively. These results show that the improved model significantly improves the feature extraction and segmentation ability of wheat leaf blight. (2) Based on the segmentation accuracy of wheat leaf blight, our improved model MSDP-SAM2-UNet segmentation accuracy reached 81.63%, which exceeded 18.58%, 19.84%, 18.4%, 20.38%, 17.49%, 16.55 and 15.52%, respectively, compared with U-Net, MultiResUnet, Swin-Unet, SMESwinUnet, UCTransNet, ACC-UNet, and SAM2.

However, this research still faces several challenges. The existing training data inadequately capture real-world environmental complexity, such as weather shifts, light intensity changes, and leaf occlusion, which can result in texture misclassification and minor lesion

omission during segmentation. Moreover, the current method cannot assess wheat leaf disease severity, which is essential for effective disease control.

We are advancing research into lightweight models for the model's field deployment, and creating a wheat leaf disease severity grading system. These efforts are crucial for automated disease detection and early prevention. Future research should enhance training data diversity, focus more on complex environments, and strengthen the model's generalization capacity.

5. Conclusions

In this study, we propose the MSDP-SAM2-UNet model for wheat leaf disease image segmentation. Our approach incorporates the RPMS module and DualPath module, significantly enhancing the network's training accuracy and achieving more precise segmentation results. Specifically, we introduce the RPMS module to improve the RFB module, expanding the receptive field and aggregating contextual information. By fusing multi-scale features, it enhances feature extraction capabilities in complex lesion areas. Additionally, residual connections prevent gradient vanishing, accelerate network training, and ensure complete information propagation. Additionally, we replace the original decoder with DualPathdecode. This approach captures irregular changes in leaf lesion edges through flexible upsampling, enabling the model to better capture edge information and preserve fine-grained features. Consequently, the model produces more refined and accurate segmentation results.

The experimental validation results demonstrate that MSDP-SAM2-UNet achieves significant improvements in PA, MPA, MIoU, FWIoU, and Dice Precision metrics, reaching 94.02%, 88.44%, 82.43%, 90.73%, 81.76%, and 81.63%, respectively. Compared to traditional methods like U-Net, our approach effectively extracts diseased regions in wheat leaves, enabling more accurate and efficient lesion segmentation.

In summary, the MSDP-SAM2-UNet model demonstrates outstanding segmentation performance for wheat leaf diseases, revealing significant potential for applications. Our approach provides a fresh perspective for image segmentation of wheat leaf diseases.

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

Funding: This research was funded by the Hubei Provincial Technological Innovation Major Project (2018A01038).

Data Availability Statement: Publicly available datasets were analyzed in this study. Details of these data can be found in the paper, which can be found at <https://arxiv.org/abs/2409.04038> (accessed on 15 October 2024). These data can be downloaded from <https://doi.org/10.5281/zenodo.13293891> (accessed on 15 October 2024).

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

Abbreviations

The following abbreviations are used in this manuscript:

SAM	Segment Anything Model
RPMS	ResPathWithMultiScale
CARAFE	Content-Aware ReAssembly of FEatures
RFB	Receptive Field Block
MLP	Multilayer Perceptron

References

1. Anam, S.; Fitriah, Z. Early blight disease segmentation on tomato plant using K-means algorithm with swarm intelligence-based algorithm. *Int. J. Math. Comput. Sci.* **2021**, *16*, 1217–1228.
2. Bakar, M.A.; Abdullah, A.; Rahim, N.A.; Yazid, H.; Misman, S.; Masnan, M. Rice leaf blast disease detection using multi-level colour image thresholding. *J. Telecommun. Electron. Comput. Eng. JTEC* **2018**, *10*, 1–6.
3. Jun, P.; Bai, Z.y.; Lai, J.-c.; Li, S.k. Automatic segmentation of crop leaf spot disease images by integrating local threshold and seeded region growing. In Proceedings of the 2011 International Conference on Image Analysis and Signal Processing, Wuhan, China, 21–23 October 2011; pp. 590–594. [\[CrossRef\]](#)
4. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical Image Analysis using Convolutional Neural Networks: A Review. *J. Med. Syst.* **2018**, *42*, 226. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–15 June 2015; pp. 3431–3440. [\[CrossRef\]](#)
6. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
7. Xu, C.; Yu, C.; Zhang, S. Lightweight multi-scale dilated U-Net for crop disease leaf image segmentation. *Electronics* **2022**, *11*, 3947. [\[CrossRef\]](#)
8. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026. [\[CrossRef\]](#)
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
10. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 8815–8827. [\[CrossRef\]](#)
11. Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. Segment anything model for medical images? *Med. Image Anal.* **2024**, *92*, 103061. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Li, Y.; Jing, B.; Li, Z.; Wang, J.; Zhang, Y. nnsam: Plug-and-play segment anything model improves nnunet performance. *arXiv* **2023**, arXiv:2309.16967. [\[CrossRef\]](#)
13. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L. Sam 2: Segment anything in images and videos. *arXiv* **2024**, arXiv:2408.00714. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Li, Z.; Tang, W.; Gao, S.; Wang, Y.; Wang, S. Adapting SAM2 Model from Natural Images for Tooth Segmentation in Dental Panoramic X-Ray Images. *Entropy* **2024**, *26*, 1059. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Xiong, X.; Wu, Z.; Tan, S.; Li, W.; Tang, F.; Chen, Y.; Li, S.; Ma, J.; Li, G. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv* **2024**, arXiv:2408.08870. [\[CrossRef\]](#)
16. Hounsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
17. Qiu, Z.; Hu, Y.; Li, H.; Liu, J. Learnable ophthalmology sam. *arXiv* **2023**, arXiv:2304.13425. [\[CrossRef\]](#)
18. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
19. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; pp. 263–273.
21. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016. [\[CrossRef\]](#)
22. Lachinov, D.A. Segmentation of Thoracic Organs Using Pixel Shuffle. In Proceedings of the SegTHOR@ ISBI, Venice, Italy, 8 April 2019.
23. Wei, T.; Chen, Z.; Yu, X.; Chapman, S.; Mello, P.; Huang, Z. Plantseg: A large-scale in-the-wild dataset for plant disease segmentation. *arXiv* **2024**, arXiv:2409.04038. [\[CrossRef\]](#)
24. Wang, Z.; Wang, Y.; An, L.; Liu, J.; Liu, H. Local Transformer Network on 3D Point Cloud Semantic Segmentation. *Information* **2022**, *13*, 198. [\[CrossRef\]](#)

25. Wang, Z.; Zhao, J.; Zhang, R.; Li, Z.; Lin, Q.; Wang, X. UATNet: U-Shape Attention-Based Transformer Net for Meteorological Satellite Cloud Recognition. *Remote Sens.* **2022**, *14*, 104. [[CrossRef](#)]
26. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [[CrossRef](#)] [[PubMed](#)]
27. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218. [[CrossRef](#)]
28. Wang, Z.; Min, X.; Shi, F.; Jin, R.; Nawrin, S.S.; Yu, I.; Nagatomi, R. SMESwin Unet: Merging CNN and transformer for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; pp. 517–526. [[CrossRef](#)]
29. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; pp. 2441–2449.
30. Ibtehaz, N.; Kihara, D. ACC-UNet: A Completely Convolutional UNet Model for the 2020s. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; pp. 692–702. [[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.