VMUnet-MSADI Visual Mamba UNet Fusion Multi-Scale Attention and Detail Infusion for Unsound Corn Kernels Segmentation

Kuibin Zhao
Henan University of Technology
Qinghui Zhang
zqh131@163.com

Henan University of Technology
Chenxia Wan
Henan University of Technology
Quan Pan
Henan University of Technology
Yao Qin
Henan University of Technology

Article

Keywords: DeepLearning, VMUnet, Multi-Scale Attention, Detail Infusion, Unsound Corn Kernels, Image Segmentation

Posted Date: November 4th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5170853/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on March 29th, 2025.
See the published version at https://doi.org/10.1038/s41598-024-80977-z.

VMUnet-MSADI： Visual Mamba UNet Fusion Multi-Scale Attention and Detail Infusion for Unsound Corn Kernels Segmentation

Kuibin Zhao[1,ǂ], Qinghui Zhang[1,*,ǂ], Chenxia Wan[1], Quan Pan[1,2], Yao Qin[1]

[1]College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China
[2]School of Automation, Northwestern Polytechnical University, Xi'an 710114, China
[*]corresponding. zqh131@163.com
[ǂ]these authors contributed equally to this work

Abstract

Corn seed breeding is a global issue, and has attracted great attention in recent years. Deploying autonomous robots for corn kernel recognition and classification has great potential in terms of constructing environment friendly agriculture, and saving manpower. Existing segmentation methods that utilize U-shaped architectures typically operate by processing images in discrete pixel-based segments. This approach often overlooks the finer pixel-level structural details within these segments, leading to models that struggle to preserve the continuity of target edges effectively. In this paper, we propose a new framework for corn seed image segmentation, called VMUnet-MSADI, which aims to integrate MSADI module into the encoder and decoder of the VMUnet architecture. Our VMUnet-MSADI model benefits from self-attention computation in VMUnet and multiscale coding to efficiently model non-local dependencies and multiscale contexts to improve the segmentation quality of different images. Unlike previous Unet-based improvement schemes, the proposed VMUnet-MSADI adopts a multiscale convolutional attention module coding mechanism at the depth level and an efficient multiscale deep convolutional decoder at the spatial level to extract coarse-grained features and fine-grained features at different semantic scales and effectively avoid the loss of information at the target boundary to improve the quality and accuracy of target segmentation. In addition, we introduce a Visual State Space (VSS) block to capture a wide range of contextual information and a Detail Infusion Block (DIB) to enhance the fusion of low-level and high-level features, which further fills in the remote contextual information during the up-sampling process. Comprehensive experiments were conducted on open-source datasets and the results demonstrate that the VMUnet-MSADI model excels in the task of corn kernel segmentation. The model achieved a segmentation accuracy of 95.96%, surpassing the leading method by 0.9%. Compared to other segmentation models, our method exhibits superior performance in both accuracy and loss metrics. Extensive comparative experiments conducted on various benchmark datasets further substantiate that our approach outperforms the state-of-the-art models. Code, pre-trained models and data processing protocols are available at https://github.com/corbining/VMUnet-MSADI

Keywords: DeepLearning; VMUnet; Multi-Scale Attention; Detail Infusion; Unsound Corn Kernels; Image Segmentation

1 Introduction

Corn is an important crop and industrial raw material, and its breeding technology has been rapidly developed, resulting in a significant increase in the number of varieties. However, the chaotic situation in the seed market is mainly due to the emergence of a large number of counterfeit and substandard seeds and thus an efficient method of corn variety identification is urgently needed to maintain the market order and safeguard the stability of agricultural production. Corn is not only one of the crops with the largest planting

area and highest yield in the world, but also occupies an important position in China's food production and its quality is directly related to national food security. Usually, corn in the harvest process using mechanized threshing methods, but in the threshing drum, pulling the stalk roller, feed churning cage and other mechanical components, corn kernels will be subjected to random extrusion, shear, kneading and other external forces, these factors may lead to damage to the corn kernels and in severe cases, even a large area of broken. During storage, corn kernels are also susceptible to a combination of environmental factors such as temperature, humidity, air and moisture, leading to problems such as germination, mold and insect damage. These abnormal kernels are detrimental to corn breeding and in the context of agricultural mechanization, especially large-scale mechanized planting, supervision and harvesting face even greater challenges. Therefore, there is an urgent need to develop a real-time, automated and accurate technique and device for identifying and detecting abnormal corn kernels in order to solve the current problems faced.

Over the past decade, convolutional neural networks (CNNs) have been widely used for various types of segmentation tasks and have made significant progress in the field of image segmentation. Especially recently, Full Convolutional Networks (FCN)[1], UNet[2] and their variants have performed particularly well. These network architectures utilize an encoder-decoder structure that enhances segmentation by directly combining the high-level semantic features extracted by the encoder path with the fine-grained features provided by the decoder path through jump connections. However, the limitations of the convolutional operation in the receptive domain and the inherent generalization bias of the convolutional structure make it difficult for CNNs to capture long-term dependencies and global contexts in the image, which restricts the further improvement of segmentation accuracy. In particular, when dealing with images of corn kernels, due to the variability of corn kernels in the target image, including changes in size, texture and shape, it is often difficult for these CNN-based methods to provide effective guidance for adapting to individual differences. It is worth noting that automatic segmentation of the image is crucial in the recognition process of corn seed images as it classifies the pixels for corn seed and determines the kind of anomaly of the seed by segmenting the size, color and shape of the anomalous key regions. Various U-structured networks[3], especially UNet[2], UNet++[4], UNet3+[5] and nnU-Net[6], have become the standard techniques to achieve high-quality segmentation results. Attention mechanisms[7] have also been integrated into these models to enhance feature mapping and improve pixel-level classification. Although attention-based models have shown improved performance, they still face significant challenges due to the high computational cost of the convolutional blocks typically used in conjunction with attention mechanisms.

Recently, visual Transformers[8] have shown remarkable potential in image segmentation tasks[9-15], mainly thanks to the Self-Attention mechanism to capture remote dependencies between pixels. To further enhance image segmentation performance, hierarchical visual transformers such as Swin[16], PVT[17], MaxViT[18], MERIT[19], ConvFormer[20] and MetaFormer[21] have been introduced. However, while Self-Attention mechanisms excel in capturing global information, they are relatively weak in understanding local spatial context [22, 23]. To address this issue, some approaches integrate local convolutional attention mechanisms in the decoder to better capture spatial details. Despite the theoretical potential of these methods, they typically require dense prediction at the pixel level, which is computationally expensive when dealing with high-resolution images, as they often rely on costly convolutional blocks. In addition, most of these methods are fixed-scale, making it difficult to cope with the diversity of scenes. Another challenge is that these image segmentation methods usually divide the image into non-overlapping chunks and convert these chunks into vector embeddings individually at each stage. However, existing chunk segmentation methods tend to ignore the pixel-level structural information and local convex topology inside each chunk, resulting in models that cannot maintain local continuity around the chunks.

To address the above limitations, we propose VMUnet-MSADI, a novel efficient multiscale convolutional attention decoding network that incorporates a multiscale convolutional block attention

mechanism and detail infusion. Specifically, VMUnet-MSADI enhances feature mapping through efficient multiscale convolution, while integrating complex spatial relations and local attention using channel attention, spatial attention and group-gated attention mechanisms. Our main contributions can be summarized as follows:

- In this paper, we propose a novel efficient multiscale convolutional decoder: we introduce an efficient Visual Mamba UNet fused multi-scale attention mechanism and detail infusion decoder for anomalous corn kernel image segmentation; this takes full advantage of the multilevel nature of the VMUnet encoder, effectively fuses the advantages of the visual Mamba UNet architecture and enhances the functionality and flexibility of the traditional encoder-decoder architecture. The core idea is to combine VMUnet with a multi-scale attention mechanism and detail infusion structure to realize automatic control of corn seed image segmentation.
- A well-designed coding mechanism for a multiscale deep convolutional attention module is proposed: we design a Multi-scale Convolutional Attention Module (MCAM) that performs deep convolution at multiple scales to improve the feature maps generated by the visual encoder. The MCAM captures salient features at multiple scales by suppressing the irrelevant regions and its high efficiency is attributed to the application of deep convolution.
- Proposed Detail Infusion Block (DIB): evaluates the spatial and channel attention scores by computing different attention mechanisms and fully utilizes the result to generate discriminative feature representations of coarse-to-fine features at different scales of the encoder, ensuring semantic consistency between coarse and fine features.
- Improved performance: extensive experiments on the corn kernel dataset provided by GaoZhe Technology show that the proposed VMUnet-MSADI consistently outperforms the existing state-of-the-art methods, which is 0.9% higher than the leading method. it can prove the effectiveness and superiority of our approach.

The rest of the paper is organized as follows. Section 2 summarizes the work related to image segmentation, Section 3 gives the data acquisition device as well as the process and Section 4 describes our proposed VMUnet-MSADI network model. Section 5 explains our experimental setup as well as the results of our benchmarks on corn seed and non-corn seed image segmentation with comprehensive experiments and visualization. Finally, Section 6 summarizes the full paper.

2 Related Work

In this section, we first summarize the most typical CNN-based methods used in corn image segmentation and then we make an overview of the recent related works about the applications of visual coders in computer vision, especially in the field of segmentation.

2.1. Corn Image Segmentation

Image segmentation involves pixel-level classification to identify detailed structures and information within images. In the field of agriculture, significant progress has been made in applying computer vision techniques and traditional algorithms to the quality assessment of corn kernels[24]. The following summarizes the key research contributions, of QUAN et al.[25] employed image labeling techniques to process images containing numerous scattered corn kernels. They utilized a multi-scale wavelet analysis algorithm for logical assessment of the labeled images, achieving segmentation, localization and shape recognition of corn kernels. CHENG et al.[26] focused on the embryonic features of corn kernels, using color space models to exploit differences in image components across different channels for segmentation. They applied an improved morphological opening and closing operation to refine the segmented images, ultimately identifying the characteristic regions of the seed's endosperm. SHI et al.[27]] analyzed collected corn images using genetic algorithms and SPSS techniques. They assessed the ratio of white to yellow areas

of damaged kernels in the HSI color space, using this information to identify corn variety attributes. K. Ding et al.[28] distinguished between whole and broken corn kernels by employing parameters such as the continuous symmetric index, curvature index and radius index. Their approach effectively excluded broken kernels from the overall dataset. Ni B et al.[29] implemented a grading system to detect whole versus broken kernels and examined the horizontal and vertical histogram distributions of corn kernels to analyze their concave and convex characteristics. Ng H. F[30] utilized staining techniques on the damaged surfaces of broken corn seeds and applied single-grain and batch analysis methods to detect internal mechanical stress cracks. A three-layer neural network was employed to effectively identify moldy corn seeds. I. Zayas et al.[31] analyzed twelve morphological parameters and selected seven significant parameters to establish a Mahalanobis distance discriminant function for distinguishing between whole and broken kernels. These studies demonstrate the application of various image processing and analysis techniques in corn kernel quality assessment, providing valuable references for research in this domain.

Recently, deep neural networks have demonstrated substantial capabilities and notable advantages in feature extraction for target object detection[32]. The application of deep learning techniques in object classification not only enhances adaptability to detected objects but also significantly improves the accuracy of detection results, with pronounced benefits observed in grain classification tasks[33]. The following provides an overview of relevant research, LIN et al.[34] designed a multi-convolutional block feature detection network model for rice classification, achieving efficient identification of rice varieties. LIU et al.[35] utilized improved YOLO series algorithms to address issues of grain breakage and counting with precise identification and detection capabilities. KHAKI et al. [36] applied convolutional neural network-based techniques to achieve high-precision identification of damaged corn kernels under varying light conditions. ZHAO et al.[37] developed a deep learning-based model for identifying and screening high-quality soybean seeds, which improved seed screening efficiency. JIN et al.[38] proposed a soybean quality detection algorithm based on an enhanced U-Net model, increasing the accuracy of soybean quality assessment. TU et al.[39] employed a VGG16 model with transfer learning to classify the JINGKE-968 variety, achieving a test accuracy of up to 98%. LV et al.[40] trained an improved ResNet model, achieving a maximum recognition accuracy of 96.4%. These studies indicate that deep learning technologies, particularly advanced network models and transfer learning strategies, play a crucial role in enhancing the accuracy and efficiency of grain classification and detection.

2.2. Vision Encoders

CNNs have become the core of basic encoders due to their advantages in processing spatial relationships in images. For example, AlexNet[32] and VGG[41] laid a solid foundation for the development of CNNs through layer-by-layer deep convolutional feature extraction. GoogleNet [42] introduced the Inception module, which makes it more efficient to compute features at multiple scales. ResNet [43] solved the gradient vanishing problem through residual concatenation, which made it possible to train deeper networks. MobileNets[44] successfully extends CNNs to mobile devices with lightweight deeply separable convolutions and EfficientNet[45] further improves the performance of CNNs with a scalable composite extension architecture design. Although CNNs perform well in numerous visual tasks, it is often difficult to capture remote dependencies in images due to the limitation of their local receptive domain.

Recently, Vision Transformers (ViTs), first proposed by Dosovitskiy et al. are able to efficiently capture remote dependencies between pixels through the Self-Attention (SA) mechanism. With the continuous development of ViTs, these models have achieved significant performance improvements by combining CNN features[18, 46], improving Self-Attention blocks and introducing innovative architectural designs[17, 47]. For example, Swin Transformer[16] introduces a sliding window attention mechanism, SegFormer[47] implements a hierarchical structure through Mix-FFN blocks, PVT[17] employs a spatially-approximate attention mechanism and PVTv2[7] is optimized with overlapping patch embedding and linear

complexity attention layers. maxViT[18] introduced multi-axis self-attention in the encoder to form a hierarchical CNN-transformer structure. These advances further advance the application and performance of visual transformers in various vision tasks.

Although visual transformers (ViTs) perform well in capturing remote pixel relationships, they still face certain challenges in capturing local spatial relationships between pixels. To address this problem, this paper proposes a multiscale attention mechanism decoder based on visual Mamba UNet. The decoder utilizes a multiscale convolutional attention module to refine the feature mapping and effectively incorporates a local DIB, which enhances the ability to model local spatial relationships. The accurate capture of local details in image segmentation tasks is further enhanced.

In summary, while deep learning technologies have been extensively applied to the evaluation of crop quality, research specifically focusing on the classification of anomalous corn kernels remains relatively limited. Previous studies have addressed the problem in low-density and relatively ideal conditions, but there remains significant room for improvement. Our research team proposes an innovative approach that utilizes the Visual Mamba UNet network, incorporating a multi-scale attention mechanism to enhance the capture of features across different scales and achieve fine-grained segmentation of corn kernels. We introduce the VSS module to capture contextual information, thereby improving adaptability to complex backgrounds and environmental variations. Additionally, we integrate the DI Block to enhance the fusion of low-level and high-level features, thereby improving the detail and accuracy of feature representation. Through these novel techniques, we aim to achieve more accurate and effective recognition of corn kernels, providing valuable insights and new directions for research in corn kernel classification.

3 Materials

3.1 Sample Preparation

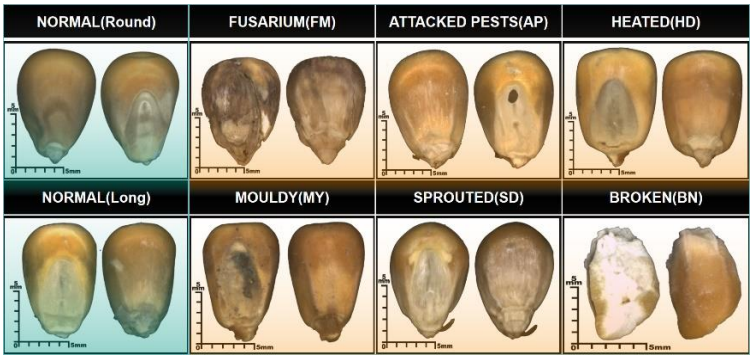


Fig 1. Examples of NOR and DU corn grains.

In this study, the data set used to train the network model was mainly from Anhui GaoZhe Technology Co., LTD.[48]. There are three kinds of data collection devices: P600, G600 and M600. The P600 consists of two industrial cameras with a light source and a conveyor belt that feeds the grain automatically. The G600 consists of an industrial camera with a light source and a conveyor belt; The M600 consists of a vision terminal and a fixed bracket. Among them, the experimental data of this project are collected by P600 and G600, which are composed of industrial cameras, grain placement platforms and lighting sources respectively. The corn grains in the data set were mainly from five countries and regions, including the United States (USA), Canada (CAN), Australia (AU), Cambodia (KHM) and China (CHN). Corn seeds can be divided into two categories, NORMAL (NOR) seeds and abnormal (Damaged and Unsound) seeds and abnormal seeds are divided into six categories, respectively, FUSARIUM & SHRIVELLED (F&S) seeds, SPROUTED (SPROUTED) seeds and damaged and unsound (DU) seeds. SD) seeds, MOULDY (MY) seeds, BROKEN (BN) seeds, ATTACKED by PESTS (AP) seeds and BLACK POINT (BP) seeds, as shown in Figure 1.

The light blue part is the normal corn grain map, including round and long grains, while the orange part shows the six abnormal grain maps. The dataset has a total of 11,460 image and the training model uses this to randomly divide the training set, verification set and test set with a ratio of about 10:1:1 to ensure their independence. The detailed distribution of selected corn kernel image data is shown in Table 1.

Table 1： Corn kernel dataset.

Corn Classes		Dataset Distribution		
		Train	**Validation**	**Test**
DU	**F&S**	1000	100	100
	SD	400	40	40
	MY	1000	300	300
	BN	1000	200	200
	AY	1000	100	100
	BP	540	120	120
NOR		4000	400	400

3.2 Images Acquisition System

The acquisition of the actual test data in this study mainly consists of the test bench of the detection camera (HIKVISION, MV-CA050-10C), the fill light strip, the image acquisition card and the computer. To ensure real-time and continuous data acquisition, we designed the bottom of the acquisition device with an adjustable Angle single-layer guide structure. It is covered with white matte stickers to reduce the effects of direct light and corn grain rolling. In the test process, corn grains start from the feeding hopper, enter the single-layer guide structure through the trough wheel and the transition plate and reach the uniform distribution plate. There is an adjustable Angle α between the plane and the horizontal plane of the uniform distribution plate and the bottom of the uniform distribution plate is equipped with a vibrator to assist in the flattening of corn grains. After passing through the uniform distribution plate, the corn grains achieve single-layer discretization and finally pass through the detection area, image acquisition is carried out by an industrial camera equipped with a 6 mm focal length adjustable lens, as shown in Figure 2.

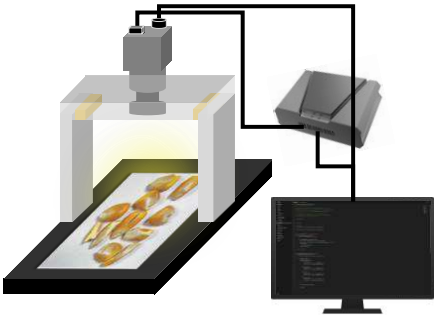


Fig 2. A schematic diagram of the hyperspectral imaging system.

At each scan for data collection, a set of 64 grains of corn from each category was removed and spread out in a tile, organized into 8 rows by 8 columns. To ensure that no two seeds come into contact, all the seeds that are placed are placed about 2 x 2 cm in the center of the platform. A total of 10 groups were designed for a test cycle with 7 types of tests and 4480 grains were processed in each test cycle.

3.3 Data Preprocessing

The data collected in the experiment took a picture containing multiple corn grains in a photo, but it was a single corn grain that was finally identified as abnormal. Therefore, the image needs to be processed. The main steps include binary image processing of multiple grains and then contour detection, calculating the

maximum external polygon according to the detected contour pixels and extracting the outline of a single corn grain according to the maximum and minimum value of the external polygon. Finally, the selected corn grains were cut to get the single corn image. The image processing process is shown in Figure 3.

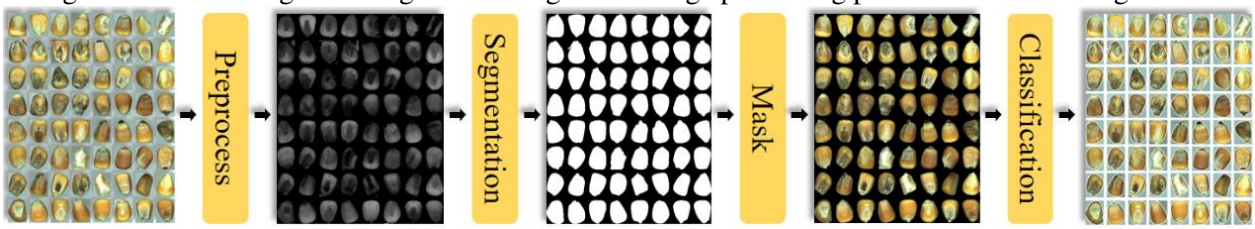
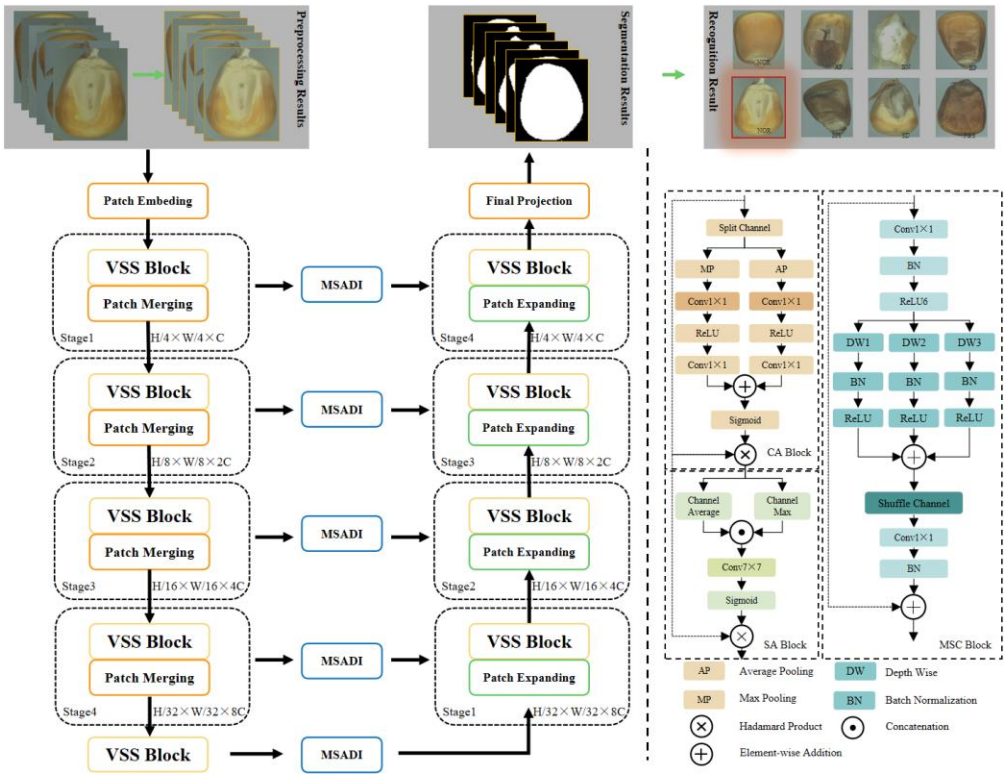


Fig 3. Corn kernel image processing process.

In the study, we used a photograph of multiple corn grains as data, and we needed to process these images in order to identify abnormal kernels in individual corn grains. The main steps of processing are as follows: First, the original image is filtered and enhanced, and CIELAB channel separation is carried out. The analysis shows that the B-channel image has high contrast, which is conducive to image segmentation. Gamma transformation is performed on the B-channel image, where $\lambda = 0.16$ and then binarization processing is carried out to obtain the binarization image of a single grain. Then, the contour detection was carried out, and the maximum external polygon of each corn kernel was calculated by counting contour pixels. Then, the outer rectangle of each corn kernel is extracted according to the maximum and minimum values of these outer polygons. Finally, each corn grain is cut according to the external rectangle, and the image of a single corn is obtained. Figure 3 shows the image after the combination of 8×8 corn grains. The whole process of image processing is shown in Figure 3. After the above processing, the isolated single corn kernel image was input into the VMUnet-MSADI network model as the actual test data for the classification and recognition of corn grains.

3.4 Data Flow



(a) (b)

Fig 4. The overall architecture of the VMUnet-MSADI model, consists of an Encoder module, MSADI module, and Decoder module. (a) The overall architecture of VMUnet-MSADI. (b) The core processing units of MSADI consist of the CA Block, SA Block and MSC Block.

The core framework of the VMUnet-MSADI model is proposed in this study. The data is preprocessed before passing through the network, including image filtering and denoising, scale transformation, color correction, normalization, single grain cutting, etc. Then feed into the network model, through the encoder stage, single image feature extraction; Then we enter the visual state space module (VSS) to capture the context information in the wide area. Then we enter the Detail Infusion Block to enhance the low-level feature detail information and the high-level feature detail information infusion for the network model. Before it enters the main 2D Selection Scan (SS2D) module, the VSS module results are output through hierarchical feature mapping combined with outputs from other layer information streams. Finally, the final recognition result is output through the decoder of the VSS module, as shown in Figure 4.

4 Methods

4.1 Preliminaries

SSM-based VMUnet-MSADI networks, including the structured state space sequence model and Mamba model, rely on a classical continuous system, represent as $x(t) \in \mathrm{R}$, which maps one-dimensional input functions or sequences to output $y(t) \in \mathrm{R}$ via an intermediate implicit state $h(t) \in \mathrm{R}^N$. This process can be expressed as a Linear Ordinary Differential Equation (ODE):

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) \end{aligned} \qquad (1)$$

Where， $A \in \mathrm{R}^{N \times N}$ represents the state matrix， $B \in \mathrm{R}^{N \times 1}$ and $C \in \mathrm{R}^{N \times 1}$ is the projection vector. Structured state space sequences and Mamba discretize this continuous system to make it more suitable for deep learning scenarios. Specifically, they introduce a time scale parameter δ and convert A and B into discrete parameters $\overline{A}$ and $\overline{B}$ using fixed discretization rules. A zero-order hold (ZOH) is usually used as a discrete rule, which can be defined as:

$$\begin{aligned} \overline{A} &= \exp(\delta A) \\ \overline{B} &= (\delta A)^{-1}(\exp(\delta A) - I) \cdot \delta B \end{aligned} \qquad (2)$$

After discretization, SSM-based models can be computed either by linear recursion or global convolution, defined as equations 3 and 4, respectively.

$$\begin{aligned} h'(t) &= \overline{A}h(t) + \overline{B}x(t) \\ y(t) &= Ch(t) \end{aligned} \qquad (3)$$

$$\begin{aligned} \overline{K} &= (C\overline{B}, C\overline{A}\overline{B}, \ldots, C\overline{A}^{L-1}\overline{B}) \\ y &= x * K \end{aligned} \qquad (4)$$

Where, $\overline{K} \in \mathrm{R}^L$ represents a structured convolution kernel and L represents the length of the input sequence x .

4.2 VMUnet-MSADI Architecture

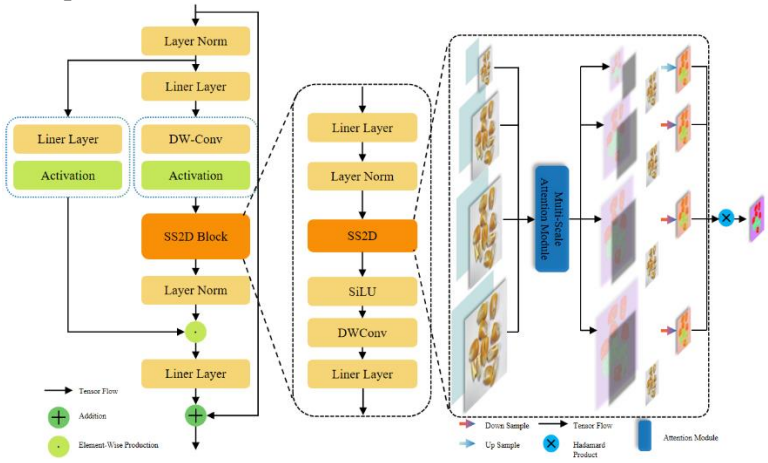
The overall structure of VMUnet-MSADI consists of three main modules: Encoder, DIB (Detail Infusion Block), and Decoder. Given the input image I , where $I \in R^{H \times W \times 3}$, the encoder generates M-levels of features. We represent the characteristics of layer i-th as f_i^o , where $1 \le i \le M$. These accumulated features as $\{f_1^o, f_2^o, ..., f_M^o\}$ are then forwarded to the DIB for further enhancement. After entering the DIB, the encoder output channel of f_i is $2^i \times C$, these accumulated features $\{f_1^o, f_2^o, ..., f_M^o\}$ enters the DIB for feature fusion, f_i

corresponds to $f_i^{'}$ as the output of i-th stage. $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2^i C$ is characterized by $f_i^{'}$. In our model, we use deep supervision to calculate the loss of $f_i^{'}$ and $f_{i-1}^{'}$ features.

In this article, we use the $[N_1, N_2, N_3, N_4]$ VSS block on the four stages of the encoder, with a channel count of $[C, 2C, 4C, 8C]$ for each stage. According to our observations in VMamba, the different values of N_3 and C are important factors in differentiating the Tiny, Small and Base framework specifications. According to the VMamba specification, we let C take the value of 96, N_1 and N_2 take the value of 2, respectively, and N_3 take the values in the set [49]. This represents our intention to use tiny and small models of VMamba as the backbone of our ablation experiments.

4.3 VSS Module

VSS Block is the backbone of VMUnet-MSADI, and SS2D is the core of VSS Block. The Detail Infusion Block consists of the trunk output feature, a multi-scale attention module, and the DIB output feature is the same size as the trunk output feature.



(a) The core structure of VSS Block. (b)Multi-Scale Attention is the core module of SS2D.

Fig 5. VMUnet-MSADI network core functional modules, VSS and MSA modules.

The VSS block, derived from VMamba, is the backbone of the VM UNetV2 encoder and its structure is shown in Figure 5(a). The input is first processed through an initial linear embedding layer, which is then split into two separate information streams. A stream passes through the 3×3 deep convolution layer and the subsequent Silu activation function before entering the main SS2D module. The SS2D output is then processed by the layer normalization layer and combined with the output of other information flows, which are also processed by Silu activation. The combined output forms the final result of the VSS block.

DIB, as shown in Figure 5(b). For hierarchical feature mapping $f_i^o = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2^i C$, the encoder generates $1 \le i \le 4$ where i represents layer i-th.

Different attention mechanisms can be used in the DIB to calculate the attention score of space and channel. According to what is mentioned in UNetV2, we use CBAM to achieve temporal and spatial attention. The calculation formula is as follows, ϕ_i^{att} represents i-th time attention calculation:

$$f_i^1 = \phi_i^{att}\left(f_i^0\right) \qquad (5)$$

We use 1×1 convolution to align the f_i^1 channels to C, and the resulting feature map is denoted as $f_i^2 \in R^{H_i \times W_i \times C}$.

DIB decoder stage i-th, f_i^2 represents the target reference. Then, we adjust the size of the feature map at each j-th layer so that it matches the size of f_i^2, as follows:

$$f_{ij}^3 = \begin{cases} \mathrm{G_d}\left(f_j^2, (H_i, W_i)\right) & if \quad j < i \\ \mathrm{G_1}\left(f_j^2\right) & if \quad j = i \\ \mathrm{G_u}\left(f_j^2, (H_i, W_i)\right) & if \quad j > i \end{cases} \qquad (6)$$

In Formula 6, Gd, Gi and Gu represent adaptive averaging pooling, identity mapping and bilinear interpolation respectively. In equation 7, θ_{ij} is the parameter of smooth convolution, and f_{ij}^4 is the j-th smooth feature map at i-th level. Here, $H()$ represents the Hadamard product. The f_i^5 is then forwarded to the decoder at layer i-th for further resolution reconstruction and segmentation.

$$\begin{aligned} f_{ij}^4 &= \theta_{ij}\left(f_{ij}^3\right) \\ f_i^5 &= H\left(\left[f_{i1}^4, f_{i2}^4, f_{i3}^4, f_{i4}^4\right]\right) \end{aligned} \qquad (7)$$

4.4 Multi-scale Attention Module (MSAM)

We introduced an efficient multi-scale convolutional attention module to refine feature maps. *MSAM* sequentially incorporates a channel attention block $CA(\bullet)$ that emphasizes channel correlations, a spatial attention block $SA(\bullet)$ designed to capture local contextual information, and an efficient multi-scale convolution block $MSC(\bullet)$ that enhances the retention of contextual relationships in feature maps. The definition of $MSAM(\bullet)$ is defined by Equation (8):

$$MSAM(x_{tensor}) = MSC(SA(CA(x_{tensor}))) \qquad (8)$$

Here, x_{tensor} represents the input tensor. Due to the use of deep convolution across multiple scales, our $MSAM$ module is more efficient compared to the convolutional attention module with a significant reduction in computational cost.

Multi-scale Convolution Block (MSCB): We introduce an efficient Multi-scale Convolution Block to enhance the features generated by our cascade expansion path. In our MSCB, we follow the design principles of the Inverted Residual Block (IRB) from MobileNetV2. However, unlike IRB, our MSCB performs deep convolution across multiple scales and utilizes channel shuffling to shuffle channels across groups. Specifically, in our MSCB, we first use a pointwise (1×1) convolution layer $PWC1(\bullet)$ to increase the number of channels (i.e., expansion factor=2), followed by batch normalization $BN(\bullet)$ and activation $ReLU6(\bullet)$. Next, we apply multi-scale depth-wise convolution $MSDC(\bullet)$ to capture context information at various scales and resolutions. Since depth-wise convolution ignores inter-channel relationships, we implement channel shuffling to facilitate cross-channel mixing. Subsequently, another pointwise convolution layer $PWC2(\bullet)$ is used, followed by $BN(\bullet)$ to revert to the original number of channels, thereby capturing channel dependencies. The definition of $MSCB(\bullet)$ is provided by Equation (9):

$$\begin{aligned} MSDC(x_{tensor}) &= ReLU6(BN(PWC1(x_{tensor}))) \\ MSCB(x_{tensor}) &= BN(PWC2(CS(MSDC))) \end{aligned} \qquad (9)$$

Where, the parallel operations with different kernel-sizes (KS) are defined as follows in Equation (10):

$$MSDC(x_{tensor}) = \sum_{k \in KS} DWCB_k(x_{tensor}) \qquad (10)$$

$DWCB_k(x_{tensor}) = ReLU(BN(DWC_k(x_{tensor})))$ is defined as follows. $DWC_k(\bullet)$ refers to the depth-wise convolution with kernel size k. $BN(\bullet)$ and $ReLU6(\bullet)$ denote batch normalization and $ReLU$ activation, respectively.

Additionally, our sequence $MSDC(\bullet)$ utilizes the recursively updated input x_{tensor}, where x_{tensor} maintains a residual connection with the previous $DWCB_k(\bullet)$ to enhance regularization, as defined in Equation (11):

$$x_{tensor} = x_{tensor} + DWCB_k(x_{tensor}) \quad (11)$$

Channel Attention Block (CAB): We employ the Channel Attention Block to assign different levels of importance to each channel, thereby emphasizing more relevant features while suppressing less useful ones. Essentially, the $CA(\bullet)$ block determines which feature maps to focus on (and subsequently refines them). We first apply maximum pooling $P_M(\bullet)$ and average pooling $P_A(\bullet)$ across the spatial dimensions (i.e., height and width) to extract the most prominent features from the entire feature map for each channel. Then, for each pooled feature map, we use pointwise convolution $PWC1(\bullet)$ to reduce the number of channels, followed by $ReLU$ activation. Next, another pointwise convolution $PWC2(\bullet)$ is applied to restore the original channel dimensions. We then add the two restored feature maps together and apply a Sigmoid activation σ to estimate the attention weights as adjustment factors. Finally, these weights are integrated into the input x_{tensor} using Hadamard product. The channel attention block $CA(\bullet)$ is defined as follows in Equation (12):

$$\begin{aligned} CAB(x_{tensor}) &= \sigma(CAB1(x_{tensor}) + CAB2(x_{tensor})) \otimes x_{tensor} \\ CAB1(x_{tensor}) &= PWC2(\mathrm{Re}LU6(PWC1(P_M(x_{tensor})))) \\ CAB2(x_{tensor}) &= PWC2(\mathrm{Re}LU6(PWC1(P_A(x_{tensor})))) \end{aligned} \quad (12)$$

Spatial Attention Block (SAB): We employ the Spatial Attention Block to simulate the attention mechanism of the human brain, focusing on specific regions of the input image. Essentially, the $SA(\bullet)$ module determines the focal points within the feature map. By enhancing these focal regions, the model's ability to recognize and respond to relevant spatial features is improved, which is crucial for image segmentation as the background and positioning of objects significantly influence the output. In $SA(\bullet)$, we first aggregate the maximum $C_M(\bullet)$ and average $C_A(\bullet)$ values along the channel dimension to focus on local features. Next, we use a large kernel convolution (such as a 7×7 kernel) to enhance the local contextual relationships between features. We then apply Sigmoid activation σ to compute the attention weights. Finally, these weights are applied to the input x_{tensor} using the Hadamard product, which allows for more targeted processing of the information. The spatial attention block $SA(\bullet)$ is defined as follows in Equation (13):

$$SAB(x_{tensor}) = \sigma(LKC([C_M(x_{tensor}), C_A(x_{tensor})])) \otimes x_{tensor} \quad (13)$$

4.4 Loss Function

For our corn anomalous grain image segmentation task, we mainly use the basic cross-entropy and Dice methods as loss functions because all of our dataset masks consist of two classes that are single target and background.

$$\begin{aligned} L_{\mathrm{BceDice}} &= \lambda_1 L_{\mathrm{Bce}} + \lambda_2 L_{\mathrm{Dice}} \\ L_{Bce} &= -\frac{1}{N}\sum_{N}^{1}\left[y_i log\left(\hat{y}_i\right) + \left(1-y_i\right) log\left(1-\hat{y}_i\right)\right] \\ L_{\mathrm{Dice}} &= 1 - \frac{2|X \cap Y|}{|X| + |Y|} \end{aligned} \quad (14)$$

Where, (λ_1, λ_2) is a constant, and (1,1) is usually selected as the default parameter.

5 Experiments

After studying the VMamba network operation, the image size of the dataset was adjusted to 256×256 pixels. In order to avoid overfitting, data enhancement methods are introduced, such as random flipping, random rotation and other operations. During training, the initialization parameters are as follows, the batch

size is set to 64, the optimizer's learning rate is set to 1e-3 and by using Cosine Annealing LR as a scheduler, the operation can be performed over a maximum of 50 iterations and the learning rate can be as low as 1e-5. We did a total of 50 epochs. For VMUnet-MSADI networks, the initial weights of the encoder cells are set to align with VMamba's weights. The implementation is carried out on the Ubuntu 20.04 system, using NVIDIA RTX 3060, 10GB memory, Python3.8, PyTorch 2.0.1 and CUDA11.7 development environment.

5.1 Datasets

We used an open-source data set provided by GaoZhe Tech to verify the validity of our framework. A total of 11,460 data were collected in the corn kernel dataset, including 2 major categories: normal (NOR) grains 4800, abnormal (DU) grains 6660 and abnormal was further divided into 6 subcategories, namely, wilted (F&S) grains 1200, germinated (SD) grains 480, moulded (MY) grains 1600. Broken (BN) seed 1400, Pest (AY) seed 1200 and black spot (BP) seed 780. The above data set is divided into normal classes and abnormal classes in a ratio of 7:3 as training and test sets.

Table 2: Dataset distribution.

Corn Classes		Split Ratio					
		7:3		8:2		9:1	
		Train	Test	Train	Test	Train	Test
DU	F&S	392	268	448	112	504	100
	SD	252	108	288	72	324	36
	MY	392	268	448	112	504	56
	BN	392	268	448	112	504	56
	AY	392	268	448	112	504	56
	BP	140	60	160	40	180	20
NOR		1050	450	1200	300	1300	200

This data set plays a crucial role in accurately identifying the categories of corn grains. Serves as the basis for a model-driven learning process that enables it to identify patterns and characteristics that are unique to each seed class. We collected 4480 corn kernel samples, among which 1500 were high-quality seed samples and the remaining 2800 were abnormal seed samples, which were further divided into 6 different categories, as shown in Table 2. In order to verify the validity of the model, we divided the data set into three groups according to different proportions for training verification, the segmentation ratio is 7:3, 8:2 and 9:1 and observed their performance. As can be seen from the experimental analysis, we find that the 8:2 ratio gives good results, so for further research, we prefer the 8:2 ratio training weight model.

5.2 Implementation Details

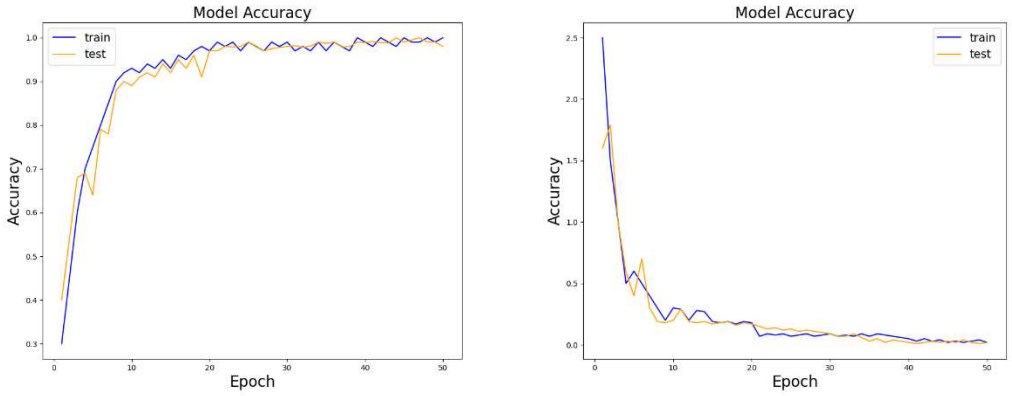


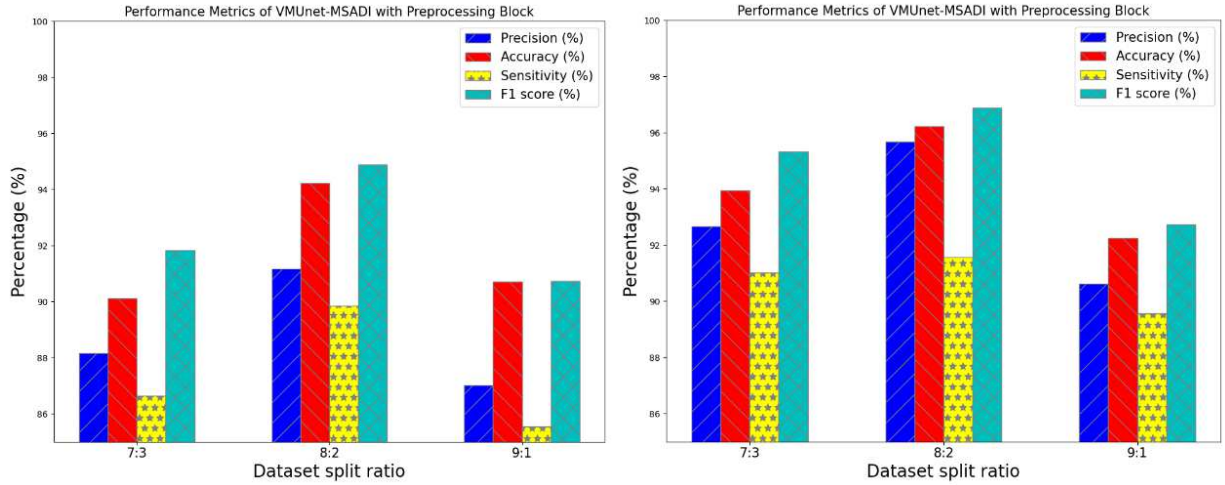
Fig 6. Graphical representation of epoch-wise accuracy and loss curve of the VMUnet-MSADI model.

In order to train an effective VMUnet-MSADI network on the corn data set, the classification cross-entropy loss function is used to reduce the influence of data set diversity and the classification cross-entropy is paired with the activation function of the final output layer. Softmax is used to convert the original model output into a probability value and the classification cross-entropy is compared with the difference between the probability value and the true value. Finally, the exact type of corn kernel was obtained. During the training process, we used the Adam optimizer, the learning rate was set to 0.0001, the weight decay was set to 0.0005, the momentum was set to 0.9 and the batch size was set to 16. In order to train the network, 50 epochs were used to output a model. Figure 6 shows that during the training and verification process, with the increase in the number of training rounds, the accuracy of the VMUnet-MSADI model increases and the loss gradually decreases.

Table 3: The results of VMUnet-MSADI with and without a preprocessing block for dataset split ratios (Bold indicates the best).

Model	*Split Ratio*	*Precision (%)*	*Accuracy (%)*	*Sensitivity (%)*	*F-1Score(%)*
VMUnet-MSADI **Without** PreProcessing	**7:3**	88.15	90.11	86.65	91.82
	8:2	**91.17**	**94.23**	**89.86**	**94.88**
	9:1	87.01	90.71	85.55	90.73
VMUnet-MSADI **With** PreProcessing	**7:3**	92.67	93.93	91.02	95.33
	8:2	**95.67**	**96.23**	**91.56**	**96.88**
	9:1	90.62	92.25	89.56	92.73

The following are four indicators: Precision (Pre), also known as Specificity (Spe), Sensitivity (Sen), also known as Recall, Accuracy (Acc) and score (F1_score). This index was used to evaluate the influence of the segmentation ratio of different data sets on the training model before and after adding the preprocessing module. Table 3 shows the results of VMUnet-MSADI for data sets 7:3, 8:2 and 9:1. First, the corn data set is passed through the preprocessing module, then the image is divided into the data segmentation ratio in Table 2 above and the network is trained.



(a) Without preprocessing block. (b) With preprocessing block.

Fig 7. Performance of VMUnet-MSADI.

The experimental results show that the data segmentation ratio is 8:2, which shows a good result. Accuracy improved by 2.5% compared to no preprocessing module. Therefore, to some extent, the

preprocessing module is helpful to improve the classification accuracy of the model. Figure 7 shows a graphical representation of VMUnet-MSADI performance with and without preprocessing module.

5.3 Main Results

In this section, we first evaluate the performance of our proposed VMUnet-MSADI framework on corn kernel image segmentation tasks using the dataset provided by Gaozhe Technology. We compare our method against state-of-the-art approaches. Additionally, we conduct ablation studies to analyze the impact of each component used in VMUnet-MSADI.

5.3.1 Experimental Settings

Baselines: In addition to the conventional UNet, our comparative experiments involve two broad categories of methods as baselines: CNN-based methods and Transformer-based methods.

- **CNN-based Methods:** Some advanced CNN-based models are introduced and compared with our proposed model VMUnet-MSADI, Model includes UNet, UNet++, Att-net, UNetV2, UTNetV2, SANet, MALUNet, VMUNet, VMUNetV2 and DoubleU-Net.
- **Transformer-based Methods:** Several Transformer-based models are considered major contenders, including TransUNet, TransFuse, Swin-unet and MCTrans.
- **VMamba-based Methods:** approaches have advantages in remote interaction modeling and linear computational complexity, including VMUNet and VMUNetV2.

Table 4: Comparative experimental results on the GaoZhe Tech Corn datasets (Bold indicates the best)

Model	*mIoU(%)*	*mDSC(%)*	*Acc(%)*	*Spe(%)*	*Sen(%)*	*Pre(%)*
Unet[2]	76.98	85.99	95.33	97.31	86.12	93.45
UnetV2[50]	80.71	89.32	94.85	96.87	88.36	92.74
Unet++[4]	78.31	87.12	94.58	95.41	88.25	93.01
Att-Unet[51]	78.43	87.12	94.21	96.25	87.05	91.65
UTNetV2[52]	77.35	87.22	95.23	98.13	84.54	90.98
SANet[53]	79.56	88.12	94.38	95.58	89.95	91.25
MALUNet[54]	80.21	89.63	94.56	96.45	89.69	92.28
VMUNet[55]	81.25	89.71	94.96	96.13	**91.12**	93.06
VMUNetV2[56]	81.38	89.65	95.06	97.16	88.96	92.53
VMUnet-MSADI	**81.47**	**89.75**	**95.96**	**97.77**	89.13	**93.61**

Implementation Details: In this study, we compare our proposed VMUnet-MSADI model with several state-of-the-art models using a dataset split of 80:20 for training and testing. To ensure a comprehensive evaluation, we include additional performance metrics beyond the traditional Precision (Pre), Specificity (Spe), Sensitivity (Sen) and Accuracy (Acc). Specifically, we also evaluate the Mean Intersection over Union (mIoU) and the Mean Dice Similarity Coefficient (mDSC). Table 4 presents the test results for different network models on the Gaofe Technology corn grain dataset. The results indicate that VMUnet-MSADI outperforms other models in terms of mIoU, DSC, Pre and Acc. Our model also exceeds the performance of the leading model, UNetV2, with an improvement of up to 5% in the mIoU metric.

5.3.2 Corn Experimental Results

Results on Corn Segmentation: This section evaluates the performance of VMUnet-MSADI for corn grain segmentation under various conditions. To ensure a fair comparison, we initially conducted segmentation experiments using one set of normal corn grains and six sets of abnormal corn grains. Additionally, we perform cross-validation across all datasets to verify the effectiveness of the proposed

VMUnet-MSADI. The comparison results with state-of-the-art methods are presented in Table 5 and the corresponding qualitative results are shown in Figure 8.

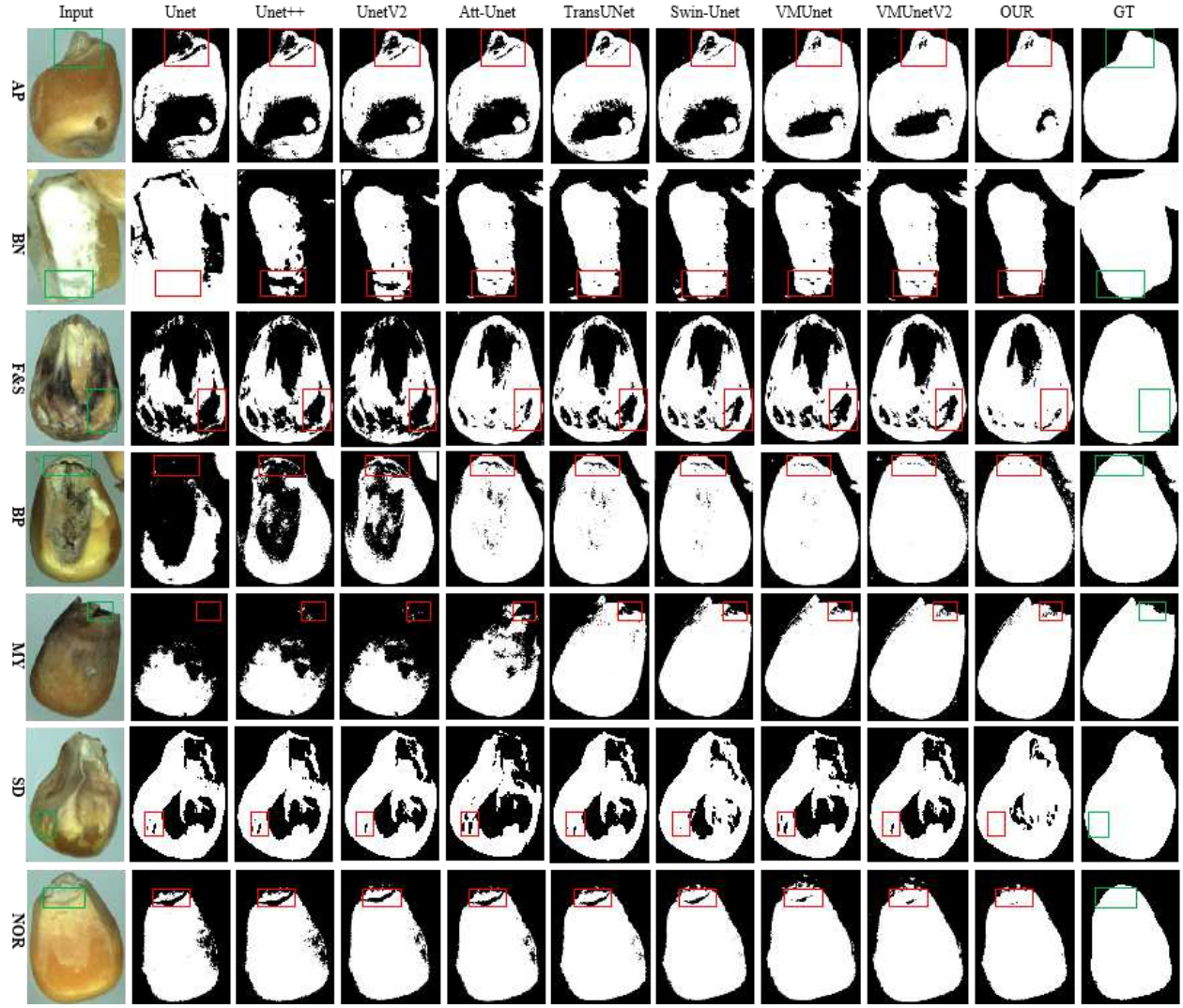


Fig 8. Comparison of qualitative results between VMUnet-MSADI and the existing models on the corn kernel segmentation task. To better visualize the differences between segmentation predictions and ground truths, we highlight the key region with appropriate boxes.

Based on the experimental results, we make the following observations. Compared to the conventional U-Net, various FCN and U-Net variants, such as VMUnet and VMUnetV2, have shown varying degrees of success. For instance, on the abnormal grain datasets, HD and MY, VMUnet and VMUnetV2 achieved increases in the mDSC scores of 0.8% and 1.6%, respectively. This underscores the significant value in further enhancing the stability and flexibility of standard encoder-decoder architectures. In contrast, Transformer-based models, such as Swin-Unet and TransUNet, have shown clear improvements over the aforementioned variants when guided by standard Transformer principles. Notably, TransUNet achieved mDSC scores of 0.881 and 0.897 on the AP and BN datasets, respectively, which are encouraging results. However, due to the limitations inherent in Transformers, these models still lag behind advanced CNN-based methods in performance. It is evident that the proposed VMUnet-MSADI achieves the highest scores across nearly all evaluation metrics for the independent datasets used.

Specifically, our VMUnet-MSADI achieved the lowest mDSC scores of 0.843 and 0.911 on the BP and FM datasets, respectively, which still represent significant improvements over previous state-of-the-art competitors such as Unet++ and Att-Unet. Moreover, Figure 8 illustrates that the network segmentation outputs from VMUnet-MSADI are more precise and detailed compared to existing baselines. This improvement not only demonstrates the effectiveness of the proposed VMUnet-MSADI but also suggests that the introduction of advanced mechanisms like multi-scale convolution and attention module offers substantial potential to surpass traditional CNN approaches. Furthermore, as shown in Table 5, the evaluation results for VMUnet-MSADI consistently outperform previous competitors across various datasets, effectively validating its generalization capability. Compared to VMUnetV2, our VMUnet-MSADI shows an increase of 2.1% in average mDSC and 2.0% in average mIoU scores. As depicted in Figure 8, VMUnet-MSADI generates high-quality segmentation masks for the corn kernel segmentation task. The promising ability of VMUnet-MSADI to identify and segment target regions of interest in corn kernel images underscores its advantages in automated corn segmentation. In summary, these comparative results confirm the superiority of the proposed VMUnet-MSADI in the automated segmentation of corn kernels.

Based on the experimental results, we have made several observations. Compared to the standard U-Net, various FCN and U-Net variants, such as VMUnet and VMUnetV2, have demonstrated varying degrees of success. For instance, on the anomaly datasets HD and MY, VMUnet and VMUnetV2 achieved increases in mDSC scores of 0.8% and 1.6%, respectively. This indicates the value of further enhancing the stability and adaptability of standard encoder-decoder architectures. In contrast, Transformer-based models such as Swin-Unet and TransUNet, guided by standard Transformer principles, significantly outperform the aforementioned variants. Notably, TransUNet achieved mDSC scores of 0.881 and 0.897 on the AP and BN datasets, respectively, which is encouraging.

Table 5: Quantitative results of corn kernel segmentation task. (Bold indicates the best)

Model	*AP*		*BN*		*FM*		*BP*		*MY*		*SD*	
	mIoU	*mDSC*	*mIoU*	*mDSC*	*mIoU*	*mDSC*	*mIoU*	*mDSC*	*mIoU*	*mDSC*	*mIoU*	*mDSC*
Unet[2]	0.769	0.829	0.709	0.823	0.701	0.791	0.702	0.725	0.722	0.765	0.712	0.785
UnetV2[50]	0.731	0.861	0.761	0.850	0.679	0.784	0.701	0.762	0.751	0.787	0.754	0.797
Unet++[4]	0.781	0.831	0.741	0.816	0.704	0.783	0.681	0.783	0.711	0.798	0.763	0.818
Att-Unet[51]	0.743	0.843	0.793	0.813	0.783	0.823	0.673	0.743	0.731	0.863	0.751	0.865
TransUNet[10]	0.811	0.881	0.816	0.897	0.781	0.867	0.681	0.786	0.726	0.836	0.766	0.828
TransFuse[15]	0.812	0.872	0.852	0.875	0.782	0.874	0.712	0.772	0.752	0.872	0.754	0.859
Swin-Unet[9]	0.821	0.891	0.831	0.912	0.801	0.871	0.721	0.821	0.771	0.891	0.778	0.864
VMUNet[55]	0.815	0.885	0.864	0.925	0.813	0.883	0.752	0.835	0.782	**0.915**	0.789	0.915
VMUNetV2[56]	0.885	0.891	0.878	0.931	**0.845**	0.910	0.785	0.836	0.785	0.907	0.795	0.916
VMUnet-MSADI	**0.889**	**0.936**	**0.891**	**0.938**	0.841	**0.911**	**0.809**	**0.843**	**0.801**	0.911	**0.821**	**0.918**

However, despite their advantages, Transformer-based models still fall short of the performance of advanced CNN-based methods due to inherent limitations. Our proposed VMUnet-MSADI has achieved the highest scores across nearly all evaluation metrics for independent datasets. Specifically, VMUnet-MSADI showed the lowest mDSC scores of 0.843 and 0.911 on the BP and FM datasets, respectively, yet these scores still surpass those of previous leading competitors such as Unet++ and Att-Unet. Furthermore, Figure 8 illustrates that VMUnet-MSADI produces more precise and detailed segmentation outputs compared to existing baselines. This improvement not only demonstrates the effectiveness of VMUnet-MSADI but also suggests that the proposed model has significant potential to surpass traditional CNN approaches. Additionally, as shown in Table 5, the evaluation results for VMUnet-MSADI consistently outperform previous competitors across various datasets, effectively demonstrating its generalization

capability. Compared to VMUnetV2, VMUnet-MSADI improves the average mDSC and average mIoU scores by 2.1% and 2.0%, respectively. Figure 9 highlights that VMUnet-MSADI generates high-quality segmentation masks for the corn kernel segmentation task, showcasing its promising ability to identify and segment areas of interest in corn kernel images. In summary, these comparative results affirm the advantages of the proposed VMUnet-MSADI in the automated segmentation of corn kernels.

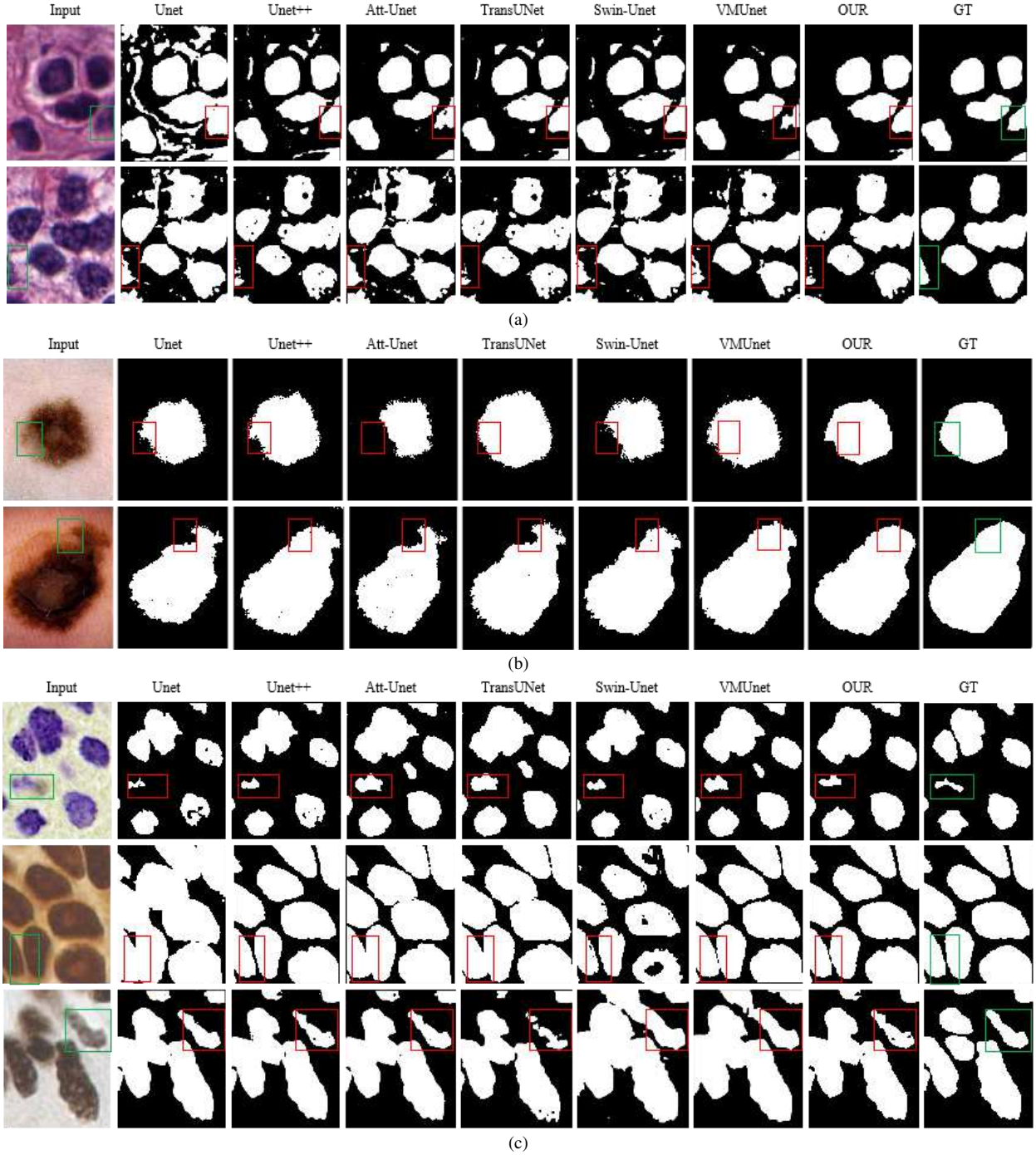


Fig 9. Qualitative results of VMUnet-MSADI on three medical image segmentation tasks compared with other models. (a) 2018 Data Science Bowl dataset (b) GLAS dataset and (c) ISIC 2018 dataset, respectively. To better visualize the differences between segmentation predictions and ground truths, we highlight the key region with appropriate boxes.

To further assess the effectiveness and generalization of the proposed VMUnet-MSADI network, we evaluated the model on three types of medical image segmentation tasks. These tasks included: skin lesion segmentation on the ISIC2018 dataset, gland segmentation on the Gland Segmentation (GLAS) dataset and nucleus segmentation on the 2018 Data Science Bowl (Bowl) dataset. To ensure fairness in the evaluation, the input image sizes were standardized to 256×256 for ISIC2018 and Bowl datasets and to 128×128 for the GLAS dataset. The ISIC2018 dataset, sourced from the ISIC-2018 challenge[11], is used for skin lesion analysis and comprises 2,596 images with corresponding annotations. In this section, we conducted experiments using five-fold cross-validation to demonstrate the efficacy of our VMUnet-MSADI model. The GLAS dataset, collected from the 2015 Histology Image Gland Segmentation Challenge, provides Hematoxylin and Eosin (H&E) stained slide images. It contains 165 images, with 85 used for training and 80 for testing, as detailed in[36]. The Bowl dataset, part of the 2018 Data Science Bowl Challenge[13], is used for nucleus detection in diverging images and consists of 670 images. We followed the same setup as[26], allocating 80% of the dataset for training, 10% for validation and 10% for testing. The experimental results are illustrated in Figure 9.

Results on 2018 Data Science Bowl: We evaluated the proposed VMUnet-MSADI network model on the nucleus segmentation task of the 2018 Data Science Bowl dataset. The comparative results with state-of-the-art methods are summarized in Table 6 and the corresponding quantitative results are shown in Figure 9(a). From Table 6, it is evident that the VMUnet-MSADI outperforms existing baselines in terms of self-attention computation and multi-scale context exploration. Specifically, our VMUnet-MSADI achieved the highest F1 score of 0.923 and a recall rate of 0.938. When compared to previous advanced methods such as TransUNet (F1 score of 0.918) and VMUNet (F1 score of 0.921), our VMUnet-MSADI shows improvements of 0.5% and 0.2%, respectively. As illustrated in Figure 9 (a), the VMUnet-MSADI model provides significantly more accurate predictions of multiple nucleus boundaries compared to existing baselines. This demonstrates the strong nucleus segmentation capability of our VMUnet-MSADI model even in challenging diverging images. These experimental results further validate the generalization capability of VMUnet-MSADI across various medical image segmentation tasks.

From Table 7, we observe the following, Attention-guided models such as U-Net (F1 score of 0.786) and U-NetV2 (F1 score of 0.812) demonstrate improvements over traditional U-Net by incorporating additional attention mechanisms. This suggests that attention can optimize the traditional U-Net model effectively. CNN-based methods that utilize multi-scale context to enhance the U-Net structure, such as Att-Unet (F1 score of 0.813) and DoubleU-Net (F1 score of 0.836), confirm the effectiveness of multi-scale context fusion. In contrast, Transformer-based models, including Swin-Unet (F1 score of 0.821), TransUNet (F1 score of 0.815) and TransFuse (F1 score of 0.818), show superior performance compared to the aforementioned methods. Our VMUnet-MSADI consistently outperforms Transformer-based competitors, improving the F1 score from 0.906 to 0.913. As shown in Figure 9(b), the VMUnet-MSADI effectively captures the boundaries of skin lesions and produces superior segmentation results. Thus, these comparative results further validate the strong capability of VMUnet-MSADI in skin lesion segmentation.

Results on GLAS Dataset: We also evaluated the VMUnet-MSADI on the GLAS dataset for gland segmentation, focusing on automatic quantification of gland morphology. The results of this comparison with state-of-the-art methods are shown in Table 8 and the quantitative results are depicted in Figure 9(c). From Table 8, we observe the following, Performance of VMUnet-MSADI: The VMUnet-MSADI obtain the highest mDSC and mIoU scores of 0.919 and 0.853. This demonstrates the effectiveness of VMUnet-MSADI in gland segmentation tasks. Comparison with State-of-the-Art Models: Compared to the previous state-of-the-art model, VMUNetV2, our VMUnet-MSADI exceeds VMUNetV2 by 14% in mDSC and 7% in mIoU scores. This significant improvement highlights the model's ability to deliver high-quality segmentation even with a limited number of training samples. The VMUnet-MSADI also shows clear advantages over recent Transformer-based approaches, including TransFuse (mDSC of 0.806), TransUNet

(mDSC of 0.801) and Swin-Unet (mDSC of 0.811). This further affirms the superiority of the dual-scale encoding mechanism and the TIF module in gland segmentation. In Figure 9(c), the visualizations of the generated masks illustrate how VMUnet-MSADI effectively differentiates between the glands and surrounding tissue. This performance highlights the model's capability to produce accurate and detailed segmentation results. Overall, these results confirm that VMUnet-MSADI delivers outstanding performance in gland segmentation, outperforming both CNN-based and Transformer-based methods.

Table 6: Quantitative results of the 2018 Data Science Bowl. (Bold indicates the best)

Model	*year*	*mIoU*	*mDSC*	*Acc*	*Spe*	*Sen*	*Pre*
Unet[2]	2015	0.911	0.756	0.712	0.803	0.719	0.754
UnetV2[50]	2018	**0.924**	0.891	0.721	0.856	0.677	0.786
Unet++[4]	2019	0.853	0.908	0.746	0.806	0.714	0.783
Att-Unet[51]	2018	0.843	0.907	0.783	0.833	0.753	0.813
TransUNet[10]	2021	0.908	0.912	0.806	0.907	0.771	0.860
TransFuse[15]	2021	0.902	0.916	0.892	0.895	0.792	0.884
Swin-Unet[9]	2021	0.851	0.915	0.861	0.912	0.811	0.879
DoubleU-Net[57]	2020	0.895	0.918	0.875	0.909	0.821	0.887
VMUNet[55]	2024	0.885	0.921	0.869	0.935	0.833	0.893
VMUNetV2[56]	2024	0.857	0.920	0.883	0.929	0.835	0.917
VMUnet-MSADI		0.886	**0.923**	**0.891**	**0. 938**	**0.841**	**0.918**

Table 7: Quantitative results of the ISIC 2018 Dataset. (Bold indicates the best)

Model	*year*	*mIoU*	*mDSC*	*Acc*	*Spe*	*Sen*	*Pre*
Unet[2]	2015	0.661	0.786	0.709	0.723	0.711	0.792
UnetV2[50]	2018	0.646	0.794	0.761	0.721	0.713	0.775
Unet++[4]	2019	0.681	0.812	0.741	0.716	0.714	0.763
Att-Unet[51]	2018	0.693	0.813	0.793	0.734	0.723	0.782
TransUNet[10]	2021	0.702	0.815	0.816	0.757	0.733	0.787
TransFuse[15]	2021	0.706	0.818	0.852	0.775	0.742	0.784
Swin-Unet[9]	2021	0.721	0.821	0.831	0.782	0.751	0.789
DoubleU-Net[57]	2020	0.735	0.836	0.852	0.796	0.765	0.802
VMUNet[55]	2024	0.745	0.845	0.864	0.805	**0.831**	0.823
VMUNetV2[56]	2024	0.785	0.875	0.878	0.831	0.815	0.834
VMUnet-MSADI		**0.793**	**0.879**	**0.895**	**0.878**	0.823	**0.852**

Table 8: Quantitative results of the GLAS Dataset. (Bold indicates the best)

Model	*year*	*mIoU*	*mDSC*	*Acc*	*Spe*	*Sen*	*Pre*
Unet[2]	2015	0.554	0.754	0.702	0.818	0.711	0.754
UnetV2[50]	2018	0.537	0.732	0.752	0.824	0.689	0.734
Unet++[4]	2019	0.575	0.751	0.735	0.816	0.714	0.773
Att-Unet[51]	2018	0.604	0.764	0.786	0.813	0.753	0.813
TransUNet[10]	2021	0.611	0.801	0.803	0.837	0.768	0.827
TransFuse[15]	2021	0.814	0.806	0.842	0.864	0.752	0.864
Swin-Unet[9]	2021	0.822	0.811	0.811	0.886	0.781	0.861
DoubleU-Net[57]	2020	0.835	0.832	0.852	0.891	0.792	0.887
VMUNet[55]	2024	0.845	0.865	0.866	0.885	0.813	0.893
VMUNetV2[56]	2024	0.846	0.905	0.879	**0.893**	0.808	0.905
VMUnet-MSADI		**0.853**	**0.919**	**0.901**	0. 892	**0.821**	**0.908**

5.4 Ablation Studies

In this section, we further conduct an ablation study of the corn kernel image segmentation task to evaluate the capabilities of each module in the proposed VMUnet-MSADI. The experimental results are shown in Table 9. Here, U-Net is considered a common baseline. "U-T" represents a U-shaped model based on the transformer encoder. "U-S" represents a pure transformer encoder similar to a U-shaped model and "U-V" represents an asymmetric encoder-decoder structure based on a U-shaped model. "U-V2" represents an asymmetric encoder-decoder structure that adopts multi-level features that introduce semantics and details. "U-MSADI" is the complete VMUnet-MSADI architecture, which contains the proposed DIB, multi-scale convolution attention module and decoder module of multi-scale attention mechanism. From Table 9, we have the following observations.

Table 9: Quantitative results of corn kernel segmentation task. (Bold indicates the best)

Model	*AP*		*BN*		*BP*		*MY*		*Average*	
	mIoU	***mDSC***	***mIoU***	***mDSC***	***mIoU***	***mDSC***	***mIoU***	***mDSC***	***mIoU***	***mDSC***
U-Net	0.769	0.829	0.709	0.823	0.702	0.725	0.722	0.765	0.725	0.785
U-T	0.731	0.861	0.761	0.850	0.701	0.762	0.751	0.787	0.736	0.815
U-S	0.781	0.831	0.741	0.816	0.681	0.783	0.711	0.798	0.728	0.807
U-V	0.815	0.885	0.864	0.925	0.752	0.835	0.782	**0.915**	0.803	0.897
U-V2	0.885	0.891	0.878	0.931	0.785	0.836	0.785	0.907	0.833	0.891
U-MSADI	**0.889**	**0.936**	**0.891**	**0. 938**	**0.809**	**0.843**	**0.801**	0.911	**0.850**	**0.906**

When we replaced the traditional encoder with a transformer-based encoder, the "U-T" achieved significant improvements in average mDSC and mIou scores of 30% and 11%, respectively. Compared with ordinary U-Net, it is proved that the transformer is good for encoding context information. In contrast, "U-T" outperforms "U-S", in particular by 8% on the average mDSC score, which validates the advantage of transformer over swin in encoders. At the same time, the decoder proposed in this paper can effectively use swin to model the remote dependency relationship, so the effect of "U-V" h and "U-V2" is not equal. By adding semantic and detail-injected asymmetric decoders, the average mIoU score of "U-V2" was improved by 30%, respectively and conversely, the average mDSC was reduced by 6%. Although the average mDSC score of "U-V2" is not as good as that of "U-V", "U-MSADI" can improve the average mDSC and mIoU scores from 0.891 to 0.906 and 0.833 to 0.850. It undoubtedly proves that MSADI module can guarantee consistency between different features and improve the segmentation performance.

Table 10: Performance comparison of model size (Params) between VMUnet-MSADI and other leading methods on the GaoZhe dataset.

Model	*year*	*mIoU*	*mDSC*	*Params*
Unet[2]	2015	0.554	0.754	24.55M
UnetV2[50]	2018	0.537	0.732	25.12M
Unet++[4]	2019	0.575	0.751	26.04M
Att-Unet[51]	2020	0.835	0.832	29.32M
TransUNet[10]	2018	0.604	0.764	25.23M
TransFuse[15]	2021	0.611	0.801	106.87M
Swin-Unet[9]	2021	0.814	0.806	114.65M
DoubleU-Net[57]	2021	0.822	0.811	148.96M
VMUNet[55]	2024	0.845	0.865	86.54M
VMUNetV2[56]	2024	0.846	0.905	96.81M
VMUnet-MSADI		0.853	0.919	103.23M

It can be seen from the above experimental results that all the designed components play an indispensable role in corn kernel image segmentation and perform well in medical image segmentation tasks. To compare model size and computational complexity, we further conducted experiments on corn kernel data sets. From Table 10, we can observe that; To expand the acceptance field, CNN-based methods usually need to stack deep enough convolutional layers, which leads to high computational costs; The self-attention mechanism requires more parameters than the convolution operation, which makes the transformer-based method larger in scale. VMUnet-MSADI can not only trade-off good complexity parameters but also obtain the best segmentation performance.

6 Conclusions

In this paper, we presented the multi-scale attention mechanism and detail infusion VMUnet (VMUnet-MSADI), a U-shaped encoder-decoder-based framework for improving the segmentation quality of corn images. Our VMUnet-MSADI was designed based on the VSS module. Besides the encoder, we also innovatively added a multi-scale deep convolutional attention module to the decoder, allowing deep convolution to be performed at multiple scales to improve the feature maps generated by the visual encoder by suppressing uncorrelated regions to capture multi-scale salient features. Moreover, we introduced a novel fused multi-scale attention mechanism within the encoder to extract features at multiple levels. We further proposed a novel DI block that leverages multi-scale attention mechanisms to evaluate spatial and channel attention scores. This module generates discriminative feature representations of both coarse and fine features across different scales, thereby ensuring semantic consistency between these features and enhancing the overall effectiveness of the encoder. Extensive experiments on corn image segmentation tasks rigorously evaluated on GaoZhe Tech's corn dataset demonstrated that our VMUnet-MSADI significantly out-performed the previous state-of-the-art methods, achieving a segmentation accuracy of 95.96%, which is 0.9% higher than the leading method. Furthermore, the inclusion of the preprocessing module enhanced the segmentation accuracy to 96.23%, marking an additional improvement of 0.27%. These results underscore the high competitiveness of our model in the segmentation task. In the future, we will focus on designing more lightweight VMUnet-based models and enhancing the model's capability to learn pixel-level texture structure features. These advancements aim to extend the generalization ability of the VMUnet architecture, improving its performance across a wider range of applications and datasets.

Acknowledgements

Zhongyuan Science and Technology Innovation Leading Talent Project (244200510024).

Author contributions

K.Z. interpretation of data, conducted the experiment performed, statistical analysis, figure generation and writing of manuscript; Q.Z. and C.W design of the VMUnet-MSADI study, and revision of the manuscript. Q.P. and Y.Q. revision of the manuscript; All authors reviewed the manuscript.

Data availability

The data that support the findings of this study are openly available in [VMUnet-MSADI] at [https://github.com/corbining/VMUnet-MSADI].

References:

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2015 [C].

[2] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation; proceedings of the Medical image computing and computer-assisted intervention – MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, F, 2015 [C]. Springer.
[3] FAN D-P, JI G-P, ZHOU T, et al. Pranet: Parallel reverse attention network for polyp segmentation; proceedings of the International conference on medical image computing and computer-assisted intervention, F, 2020 [C]. Springer.
[4] ZHOU Z, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. Unet++: A nested u-net architecture for medical image segmentation; proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, F, 2018 [C]. Springer.
[5] HUANG H, LIN L, TONG R, et al. Unet 3+: A full-scale connected unet for medical image segmentation; proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), F, 2020 [C]. IEEE.
[6] ISENSEE F, JAEGER P F, KOHL S A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation [J]. Nature methods, 2021, 18(2): 203-11.
[7] DONG B, WANG W, FAN D-P, et al. Polyp-pvt: Polyp segmentation with pyramid vision transformers [J]. arXiv preprint arXiv:210806932, 2021.
[8] NEIL H, DIRK W. Transformers for Image Recognition at Scale [J]. Online: https://ai googleblog com/2020/12/transformers-for-image-recognitionat html, 2020.
[9] CAO H, WANG Y, CHEN J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation; proceedings of the European conference on computer vision, F, 2022 [C]. Springer.
[10] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation [J]. arXiv preprint arXiv:210204306, 2021.
[11] CODELLA N, ROTEMBERG V, TSCHANDL P, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic) [J]. arXiv preprint arXiv:190203368, 2019.
[12] VALANARASU J M J, SINDAGI V A, HACIHALILOGLU I, et al. Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations; proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4 – 8, 2020, Proceedings, Part IV 23, F, 2020 [C]. Springer.
[13] CAICEDO J C, GOODMAN A, KARHOHS K W, et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl [J]. Nature methods, 2019, 16(12): 1247-53.
[14] WANG J, HUANG Q, TANG F, et al. Stepwise feature fusion: Local guides global; proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, F, 2022 [C]. Springer.
[15] ZHANG Y, LIU H, HU Q. Transfuse: Fusing transformers and cnns for medical image segmentation; proceedings of the Medical image computing and computer assisted intervention – MICCAI 2021: 24th international conference, Strasbourg, France, September 27 – October 1, 2021, proceedings, Part I 24, F, 2021 [C]. Springer.
[16] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, 2021 [C].
[17] WANG W, XIE E, LI X, et al. Pvt v2: Improved baselines with pyramid vision transformer [J]. Computational Visual Media, 2022, 8(3): 415-24.
[18] TU Z, TALEBI H, ZHANG H, et al. Maxvit: Multi-axis vision transformer; proceedings of the European conference on computer vision, F, 2022 [C]. Springer.

[19] RAHMAN M M, MARCULESCU R. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation; proceedings of the Medical Imaging with Deep Learning, F, 2024 [C]. PMLR.
[20] LIN X, YAN Z, DENG X, et al. ConvFormer: Plug-and-play CNN-style transformers for improving medical image segmentation; proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, F, 2023 [C]. Springer.
[21] YU W, LUO M, ZHOU P, et al. Metaformer is actually what you need for vision; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2022 [C].
[22] CHU X, TIAN Z, ZHANG B, et al. Conditional positional encodings for vision transformers [J]. arXiv preprint arXiv:210210882, 2021.
[23] ISLAM M A, JIA S, BRUCE N D. How much position information do convolutional neural networks encode? [J]. arXiv preprint arXiv:200108248, 2020.
[24] SEN N, SHAOJIN M, YANKUN P, et al. Research Progress of Rapid Optical Detection Technology and Equipment for Grain Quality [J]. Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery, 2022, 53(11).
[25] QUAN L-Z, XIAO-YU M. Adjusting the shape of corn kernels based on wavelet analysis [J]. Journal of Agricultural Mechanization Research, Papers, 2006: 2,154-6.
[26] CHENG H, SHI Z, YIN H, et al. Detection of multi-corn kernel embryos characteristic using machine vision [J]. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(19): 145-51.
[27] SHI ZHIXING S Z, CHENG HONG C H, LI JIANGTAO L J, et al. Characteristic parameters to identify varieties of corn seeds by image processing [J]. 2008.
[28] DING K, GUNASEKARAN S. Shape feature extraction and classification of food material using computer vision [J]. Transactions of the ASAE, 1994, 37(5): 1537-45.
[29] NI B, PAULSEN M, REID J. Corn kernel crown shape identification using image processing [J]. Transactions of the ASAE, 1997, 40(3): 833-8.
[30] NG H, WILCKE W, MOREY R, et al. Machine vision evaluation of corn kernel mechanical and mold damage [J]. Transactions of the ASAE, 1998, 41(2): 415-20.
[31] CONVERSE H, STEELE J. Discrimination of whole from broken corn kernels with image analysis [J]. Transactions of the ASAE, 1990, 33(5): 1-1646.
[32] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012, 25.
[33] PANCHAL A V, PATEL S C, BAGYALAKSHMI K, et al. Image-based plant diseases detection using deep learning [J]. Materials Today: Proceedings, 2023, 80: 3500-6.
[34] LIN P, LI X, CHEN Y, et al. A deep convolutional neural network architecture for boosting image discrimination accuracy of rice species [J]. Food and Bioprocess Technology, 2018, 11: 765-73.
[35] LIU Z, WANG S. Broken corn detection based on an adjusted YOLO with focal loss [J]. IEEE Access, 2019, 7: 68281-9.
[36] KHAKI S, PHAM H, HAN Y, et al. Convolutional neural networks for image-based corn kernel detection and counting [J]. Sensors, 2020, 20(9): 2721.
[37] ZHAO G, QUAN L, LI H, et al. Real-time recognition system of soybean seed full-surface defects based on deep learning [J]. Computers and Electronics in Agriculture, 2021, 187: 106230.
[38] JIN C, LIU S, CHEN M, et al. Online quality detection of machine-harvested soybean based on improved U-Net network [J]. Trans Chinese Soc Agric Eng, 2022, 38: 70-80.
[39] TU K, WEN S, CHENG Y, et al. A non-destructive and highly efficient model for detecting the genuineness of maize variety'JINGKE 968′ using machine vision combined with deep learning [J]. Computers and Electronics in Agriculture, 2021, 182: 106002.
[40] L ü MENGQI Z. Researchon seedclassificationbasedonimprovedResNet [J]. Journal ofChinese Agricultural Mechanization, 2021, 42(4): 92-8.
[41] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:14091556, 2014.

[42] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2015 [C].
[43] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].
[44] ANDREW G, MENGLONG Z. Efficient convolutional neural networks for mobile vision applications [J]. Mobilenets, 2017, 10: 151.
[45] TAN M. Efficientnet: Rethinking model scaling for convolutional neural networks [J]. arXiv preprint arXiv:190511946, 2019.
[46] KOESHARDIANTO M, AGUSTIONO W, SETIAWAN W. Classification of corn seed quality using residual network with transfer learning weight [J]. Elinvo (Electronics, Informatics, and Vocational Education), 2023, 8(1): 137-45.
[47] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers [J]. Advances in neural information processing systems, 2021, 34: 12077-90.
[48] FAN L, DING Y, FAN D, et al. GrainSpace: A large-scale dataset for fine-grained and domain-adaptive recognition of cereal grains; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2022 [C].
[49] SONG K, ZHANG Y, SHI T, et al. Rapid detection of imperfect maize kernels based on spectral and image features fusion [J]. Journal of Food Measurement and Characterization, 2024, 18(5): 3277-86.
[50] PENG Y, SONKA M, CHEN D Z. U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation [J]. arXiv preprint arXiv:231117791, 2023.
[51] OKTAY O. Attention u-net: Learning where to look for the Pancreas [J]. arXiv preprint arXiv:180403999, 2018.
[52] GAO Y, ZHOU M, LIU D, et al. A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv 2022 [J]. arXiv preprint arXiv:220300131.
[53] WEI J, HU Y, ZHANG R, et al. Shallow attention network for polyp segmentation; proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27 – October 1, 2021, Proceedings, Part I 24, F, 2021 [C]. Springer.
[54] RUAN J, XIANG S, XIE M, et al. MALUNet: A multi-attention and light-weight unet for skin lesion segmentation; proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), F, 2022 [C]. IEEE.
[55] RUAN J, XIANG S. Vm-unet: Vision mamba unet for medical image segmentation [J]. arXiv preprint arXiv:240202491, 2024.
[56] ZHANG M, YU Y, JIN S, et al. VM-UNET-V2: rethinking vision mamba UNet for medical image segmentation; proceedings of the International Symposium on Bioinformatics Research and Applications, F, 2024 [C]. Springer.
[57] JHA D, RIEGLER M A, JOHANSEN D, et al. Doubleu-net: A deep convolutional neural network for medical image segmentation; proceedings of the 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS), F, 2020 [C]. IEEE.