AMS-MLP: Adaptive Multi-Scale MLP Network with Multi-Scale Context Relation Decoder for Pepper Leaf Segmentation

Jiangxiong Fang , Chao Jiang , Huaxiang Liu , Houtao Jiang , Youyao Fu *

Posted Date: 23 May 2024

doi: 10.20944/preprints202405.1584.v1

Keywords: Multi-scale MLP; Pepper leaf extraction; Context relation decoder; Multi-path aggregation module

Article

AMS-MLP: Adaptive Multi-Scale MLP Network with Multi-Scale Context Relation Decoder for Pepper Leaf Segmentation

Jiangxiong Fang [1], Chao Jiang [2], Huaxiang Liu [1], Houtao Jiang [2] and Youyao Fu [1,*]

[1] School of Electronic & Information Engineering, Taizhou University, Taizhou 318000 China
[2] Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang, 330013, China
* Correspondence: fuyouyao828@126.com

Abstract: Pepper leaf segmentation plays a crucial role in monitoring pepper leaf diseases in various backgrounds and ensuring the healthy growth of peppers. However, existing transformer-based segmentation methods suffer from computational inefficiency, excessive parameterization, and limited utilization of edge information. To tackle these challenges, we propose an adaptive multi-scale MLP framework, named AMS-MLP, which combines the multi-path aggregation module (MPAM) and the multi-scale context relation mask module (MCRD) to refine the object boundaries in pepper leaf segmentation. AMS-MLP consists of an encoder-based network, an adaptive multi-scale MLP (AM-MLP) module, and a decoder network. In the encoder network, the MPAM module effectively fuses five-scale features to generate a single-channel mask, improving the accuracy of pepper leaf boundary extraction. The AM-MLP module divides the input features into two branches: the global multi-scale MLP branch captures long-range dependencies between image information, while the local multi-scale MLP branch focuses on extracting local feature maps. Adaptive attention mechanism is designed to dynamically adjust the weights of global and local features. The decoder network incorporates the MCRD module into the convolutional layer, enhancing the extraction of boundary features. To verify the performance of the proposed method, we conducted extensive experiments on three pepper leaf datasets with different backgrounds. The results demonstrate mIoU scores of 97.39%, 96.91%, and 97.91%, as well as F1 scores of 98.29%, 97.86%, and 98.51%, respectively. Comparative analysis with U-Net and state-of-the-art models reveals that the proposed method dramatically improves the accuracy and efficiency of pepper leaf image segmentation.

Keywords: multi-scale MLP; pepper leaf extraction; context relation decoder; multi-path aggregation module

1. Introduction

Pepper holds a prominent position in global agriculture, with China standing as the largest producer and consumer, accounting for 37% of the total pepper planting area worldwide. As an indispensable vegetable in our daily lives, pepper plants are highly vulnerable to diseases, particularly affecting the leaves and leading to the occurrence of frontal diseases [1]. The health and productivity of pepper plants are significantly impacted when the timely detection and control of pepper leaf diseases are neglected, resulting in substantial economic losses [2]. To combat these diseases, the primary approach is the application of pesticides [3]. However, the conventional methods of agricultural chemical application often overlook the severity of the disease, leading to inconsistent dosages administered across different areas, with some areas receiving insufficient treatment while others experience excessive chemical exposure [4,5]. Moreover, the incidence of pepper leaf diseases has shown a consistent increase over time. Inaccurate application of agricultural

 doi:10.20944/preprints202405.1584.v1

chemicals not only contributes to environmental pollution but also hampers the effectiveness of disease treatment [6]. In practice, planters typically resort to manual identification of disease spots and the assessment of disease severity. However, this manual identification process is labor-intensive and prone to subjective misunderstandings, introducing a risk of misinterpretation [7]. In view of this, numerous researchers have proposed intelligent recognition methods for plant diseases, and they have used CNN models to classify plant images with good recognition results [8–13]. These plant images are categorized based on their background into two types: natural background images and pure background images. Natural background images are obtained under plants' natural growth conditions, which often result in complex backgrounds characterized by high levels of uncertainty. In contrast, pure background images are captured against a steady backdrop, such as a tabletop or the ground, facilitating easier and more reliable disease identification. However, the backgrounds of these pure background images are usually highly individualistic and they are often difficult to replicate in the field, posing a significant challenge to the robust performance of the recognition models. Therefore, accurate extraction of diseased leaves plays a crucial role in intelligent plant disease recognition, which can remarkably improve the robust performance of disease recognition models. Among various techniques, image segmentation methods provide a direct and effective means of extracting pepper leaves, thereby laying the foundation for monitoring and diagnosing areas affected by pepper leaf diseases.

Advances in imaging technology have greatly facilitated the use of image segmentation techniques for plant leaf extraction, enabling agricultural experts to analyze plant growth based on various leaf image features [14–16]. Numerous approaches have been proposed for the segmentation of plant leaf images. Threshold-based methods, such as the fuzzy C-means algorithms [17], have been used to iteratively determine the optimal threshold for leaf image segmentation. Histogram intensity-based threshold methods [18,19], employing histogram bimodal and OSTU algorithms, had also been utilized for segmenting given leaf images. However, these threshold-based approaches faced challenges when dealing with complex images. Region-based methods, such as the region-based level set method, the region growth method [20], the wavelet method [21], have demonstrated high accuracy and fast runtime for plant leaf segmentation. Although these methods have achieved satisfactory results to some extent, their effectiveness heavily relied on image features, limiting their widespread applicability. Clustering-based methods, utilizing fuzzy k-means clustering [22], have been employed to compute clustering centers for leaf segmentation. However, these methods may encounter difficulties in escaping local optima, resulting in lower segmentation accuracy.

Deep learning-based technologies have made remarkable strides in the field of computer vision, bringing about significant advancements in agriculture applications [23]. Notably, the convolutional neural network [24] and U-Net architectures [25] have gained widespread application. Afterwards, U-Net and its variants have been presented to improve the segmentation by fusing different mechanisms, such as the attention U-Net [26], UNet++ [27], and the pyramid attention network [28]. The transformer based networks (Dosovitskiy et al., 2020) utilized the self-attention mechanism to build long-term relationships of dependency and could obtain competitive results in image recognition. It was noted that the transformer-based model [29–32] had mainly focused on improving the ability to extract the global context information and ignored the detailed information. MLP-Mixer [33] showed that pure MLP-based networks could achieve competitive performance in image segmentation since MLP can replace the self-attention mechanism in some extent. The cycle MLP network (RepMLPNet) [34] used local injection to merge the local priority into the fully connected layer, which provided a solution to extract detailed information.

Inspired by MLP-based models in replacing the self-attention mechanism, we propose a novel approach called Adaptive Multi-Scale MLP (AMS-MLP) network for pepper leaf segmentation. The AMS-MLP network follows an encoder-decoder architecture by combining the Multi-Path Aggregation Mask (MPAM) module and the Multi-scale Context Relation Decoder (MCRD) module. Moreover, to facilitate the fusion of global and local information between the encoder and decoder, an Adaptive Multi-Scale Global and Local MLP (AMSGL-MLP) module is designed to replace each skip connection layer. In the AMS-MLP network, the AMSGL-MLP module overcomes the

 doi:10.20944/preprints202405.1584.v1

limitations associated with the inductive bias of convolution layers, which can deal with the global information and progressively fuse discrete local details. Additionally, the MCRD module enables our model to emphasize the border relationship between the foreground and the background, especially at segmented edges. Our contributions are as follows:

1) We present a novel segmentation framework designed specifically for accurate pepper leaf segmentation in diverse backgrounds. This framework effectively extracts pepper leaves from images containing various pure backgrounds. Notably, different from the previous framework, a five-layer aggregation feature is utilized to generate a single-channel mask for refining the boundaries of the segmented pepper leaves.

2) We propose an AMSGL-MLP module based on the self-attention mechanism for automatic extraction of multi-scale features. The AMGL-MLP module employs two branches consisting of a Global Multi-Scale MLP (GMS-MLP) branch and a Local Multi-Scale MLP (LMS-MLP) branch to extract both global and local feature maps. By utilizing an attention mechanism, the module dynamically adjusts the weights assigned to the global and local features, facilitating the effective fusion of global and local information.

3) The MCRD module is proposed, which combines adjacent scale features using an attention mechanism to generate enhanced boundary features and contextual information for the segmented target.

4) Extensive experiments conducted on the pepper leaf dataset demonstrate the superiority of the proposed model over state-of-the-art (SOTA) methods.

The remainder of the paper is organized as follows. Section 2 provides an overview of related work, encompassing CNN-based and MLP-based models in semantic segmentation. Section 3 details the specific network architecture employed in our approach. Section 4 describes the experimental settings and presents the obtained results. Finally, Section 5 presents the conclusions drawn from this research.

2. Related Works

2.1. CNN-Based Model for Semantic Segmentation

Deep learning has emerged as a powerful approach for semantic segmentation, yielding significant advancements over traditional segmentation methods. Long et al. [24] introduced fully convolutional networks (FCN) as a notable milestone in image segmentation. These networks replaced the traditional fully connected layers with specialized convolutional layers. The design of these convolutional layers was tailored to cater specifically to the requirements of image segmentation tasks. Building upon this advancement, Ronneberger et al. [25] made contributions by introducing the U-shape network (U-Net) with the encoder-decoder architecture. U-Net consisted of five-layer convolutional blocks. These blocks incorporated two 3×3 convolutions, ReLU activation, and pooling layers to downsample the input features. The expanding decoder gradually reduced the number of feature maps while increasing their size through upsampling operations. Notably, skip connections were introduced to establish connections between low-level and high-level features. Finally, a 1×1 convolutional layer was employed to classify individual pixels based on the mapped feature vectors. While U-Net has demonstrated competitive performance, researchers have continuously sought to enhance its effectiveness by incorporating various mechanisms tailored to specific tasks. For instance, Alom et al. [35] introduced a residual mechanism (R2U-Net) into U-Net to promote feature fusion. Inspired by the work [36], Jin et al. [37] proposed a deformable network that leveraged deformable convolutional blocks to replace the traditional convolutional layers. These deformable blocks effectively captured adaptive perceptual fields through the use of dense connections. Xiang et al. [38] developed a bidirectional network (BIONet) that employed recurrent bi-directional skip connections. BIONet incorporated a multi-level network with an optimization algorithm to explore more effective architectures and extract enhanced spatial information. These advancements highlight the ongoing efforts to refine and improve semantic segmentation models beyond the original U-Net architecture.

 doi:10.20944/preprints202405.1584.v1

To enhance the receptive field and improve segmentation accuracy, attention mechanisms have been integrated into segmentation networks. One such approach is the squeeze-and-excitation network (SE-Net) [39], which utilizes channel attention to capture global image information. Another method, called attention-guided network [40], focuses on suppressing irrelevant information by expanding training samples through adversarial samples during preprocessing. A parallel reverse attention network (PraNet) [41] introduced a reverse attention block to build relationships among object regions and boundaries. The network aggregated high-level features into a parallel decoder and utilizes the generated features to extract boundary cues. Nevertheless, the inverse attention approach highlights only the pixels surrounding the segmentation result, potentially leading to the retention of erroneous pixels in the final output. In the field of natural language processing (NLP), the transformer-based models [42] have achieved significant success through their self-attention mechanism. Inspired by transformers, researchers have explored the application of transformer-based models in image analysis tasks. For instance, TransUNet [29] combines the U-Net architecture with transformers to leverage high-level informative features for improved performance.

2.2. MLP-Based Model

In recent years, fusion networks incorporating MLP-based modules have emerged as promising approaches for image segmentation, as MLPs had shown potential in partially replacing the self-attention mechanism. Notably, MLP-Mixer [33] introduced a mixed token scheme to facilitate information transmission between spatial features, achieving competitive results in image classification. This approach demonstrated the efficacy of MLPs in handling certain tasks. The feed-forward layer approach was used to replace the attention layer of the Visual Transformer (ViT) model [43], whose network architecture was similar to MLP-Mixer. Experimental results had indicated that the MLP-based network could achieve comparable performance when compared to the CNN-and-Transformer models. Gated MLPs (gMLP) [44] had been utilized to replace the transformer module, demonstrating that NLP tasks could achieve comparable results even without self-attention, as seen in transformer-based models. The cycle fully connected layer-based network [45] capitalized on the advantages of fully connected channels with arbitrary input scales. By utilizing an MLP-like architecture, the network achieved linear computational complexity while effectively expanding the receptive field to improve context aggregation. The multi-axis MLP (MAXIM) [46] adopted an MLP-based approach within a U-shape network structure to capture long-range interactions. MAXIM comprised two key blocks: a multi-axis gated MLP module and a cross-gating block. The former extracted global and local spatial features, while the latter facilitated cross-feature fusion, resulting in hybrid features that integrate both global and local information. RepMLPNet [34] integrated the training parameters of parallel convolution kernels into a fully-connected layer and incorporated local injection to merge local priorities, which could effectively capture both detailed and overall information, establishing it as a fully MLP-based structure.

It is worth noting that the convolutional method has been extensively employed in vision tasks and has demonstrated excellent segmentation performance. Valanarasu et al. [47] introduced a convolutional MLP-based network called UNeXt with a U-shape architecture. The encoding stage comprised three early convolution blocks and two tokenized MLP blocks. The tokenized MLP-based blocks projected the preceding features into a sequence of tokens and utilized MLPs to capture global information, thereby enabling pixel-wise classification. The CM-MLP framework [48] emphasizes the importance of appropriately designed MLP-based modules. CM-MLP encompassed the fusion of a multi-scale feature interaction (MSFI) block and an axial context encoder (ACE) block. The MSFI block facilitated the incorporation of local information through cascaded multi-scale MLP operations. Conversely, the ACE block focused on establishing edge relations between the foreground and background regions. These recent advancements underscored the ongoing exploration and integration of MLP-based architectures within convolutional frameworks, showcasing their potential in enhancing segmentation performance and effectively capturing essential visual features.

3. Materials and Methods

 doi:10.20944/preprints202405.1584.v1

3.1. Image Datasets

The pepper leaf image datasets (PLID) utilized in this study were obtained from the Nanchang Academy of Agricultural Sciences, collected from their farm located in Nanchang city, Jiangxi Province, China. The dataset was compiled through multi-view photography conducted between morning and afternoon on August 12 and 13, 2022. The camera used for image acquisition was equipped with an F5.6 lens and an EF-S 18-135mm f/3.5-5.6 IS USM microlens manufactured by Canon Company (Japan). The dataset comprises various instances of diseased pepper leaves, as well as a hybrid database. During the pepper growth, these leaves were severely affected by several common diseases, such as early blight disease, brown spot disease, and spot disease. In the experiment, to enhance the diversity of the dataset, leaf mold(LM), healthy pepper leaves (HPL) and viral diseases (VD) were also collected and incorporated as part of the mixed database. By encompassing a wide range of diseased and healthy instances, the PLID dataset serves as a valuable resource for studying and developing methodologies related to pepper leaf segmentation and disease classification in agricultural research.

To evaluate the effectiveness of the proposed model in segmenting actual pepper leaves, we constructed four distinct pepper leaf datasets to assess both the backbone network and the proposed model. We employed an open-source tool called LabelMe to enable manual annotation of images, allowing us to build image datasets specifically for image segmentation tasks. The foreground regions were assigned an intensity value of 1, and the background regions were assigned an intensity value of 0, which facilitated the utilization of the annotated images for training and evaluation of the proposed model. These datasets, namely Early Blight Dataset (EBD), Brown Spot Dataset (BSD), and Mixed Leaf Dataset (MLD), encompassed1190, 1385, and 6613 images, respectively. Table 1 presents the distribution of the four image datasets. The MLD dataset was a combination of EBD, BSD, Viral Diseases Dataset(VDD), Leaf Mold Dataset(LMD) and Healthy Pepper Leaf Dataset(HPLD). To ensure a comprehensive evaluation, each dataset was divided into three subsets: 70% for training, 10% for validation, and 20% for testing. Figure 1 shows several representative examples from the EBD, BSD, MLD and HPLD, respectively. In the experiment, we standardized the image size for each dataset to 512×512 pixels, facilitating consistent processing and analysis across all datasets.

Table 1. The distribution of the four image datasets.

Dataset	Test	Training	Validation	Total
Early Blight Dataset (EBD)	163	865	162	1190
Brown Spot Dataset (BSD)	186	1015	184	1385
Mixed Leaf Dataset(MLD)	1323	4629	661	6613

 doi:10.20944/preprints202405.1584.v1

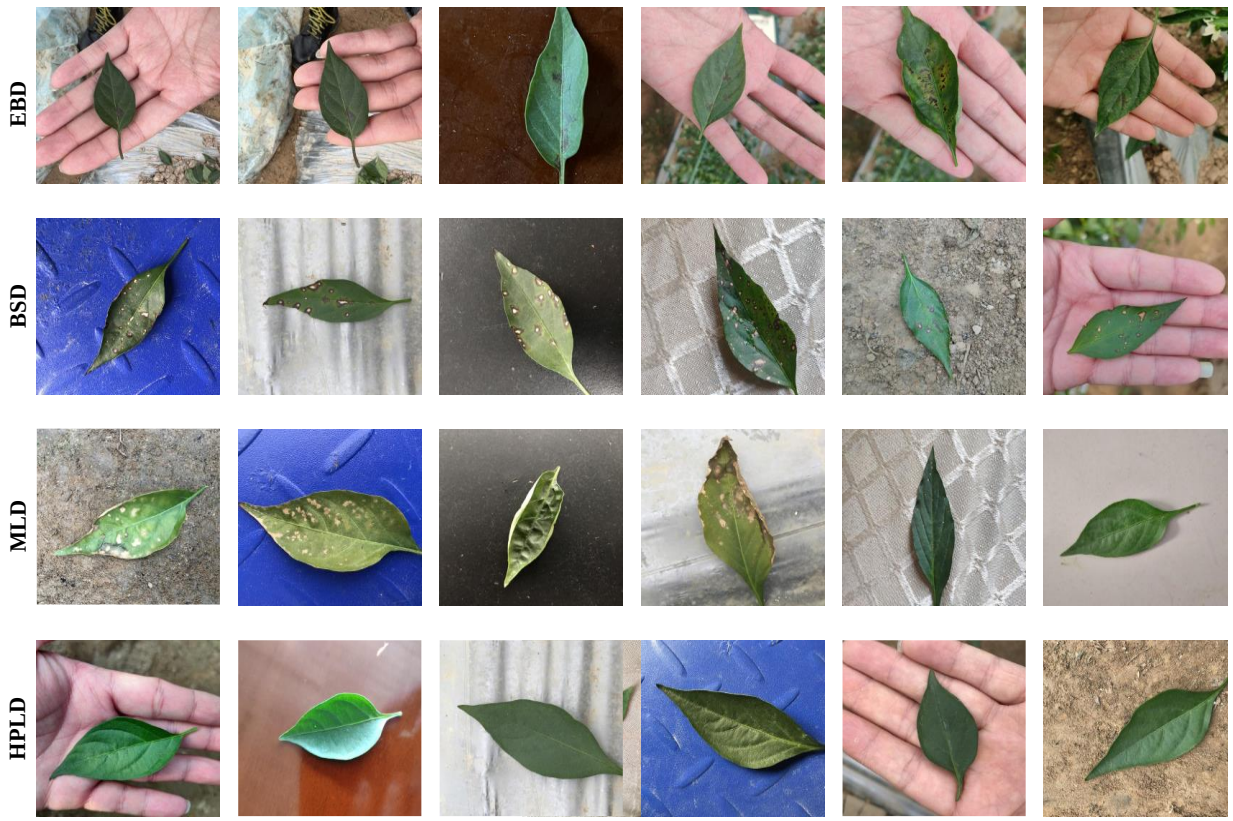

Figure 1. The sample images in different pure backgrounds.

3.2. Method

In this section, we will provide an overview of the AMS-MLP model and discuss the incorporation of three key modules within the encoder-decoder architecture. These modules consist of the adaptive multi-scale MLP module, the multi-scale context relation decoder module, and the multi-path aggregation mask module. Additionally, we will present the loss function utilized in the model. By integrating these modules and utilizing an appropriate loss function, the AMS-MLP model demonstrates improved performance in image segmentation tasks.

3.2.1. Overall Architecture

Figure 2 illustrates the network architecture of the proposed AMS-MLP network, which is based on a U-shape architecture. The AMS-MLP model is composed of three essential components: the encoder network, the adaptive multi-scale MLP (AM-MLP) module, and the decoder network. The encoder network comprises five convolutional layers, incorporating four downsampling operations and an MPAM module. Specifically, the feature map undergoes a series of five convolutional blocks to extract multi-scale deep features. Each convolutional block consists of a 3×3 convolutional layer with a step size of 1 and a padding of 1, followed by a batch normalization layer, a ReLU activation function, and a max-pooling operation with a stride of 2. Moreover, the multi-scale features at layers 1 through 5 in the encoder network are fused within the MPAM module to produce a coarse mask, which is subsequently input to the MSRD module to capture the edge information. The decoder network comprises five convolutional blocks, which incorporate four upsampling layers, and three MSRD modules. Each convolutional block in the decoder consists of a 3×3 convolutional layer, a batch normalization layer, and a ReLU activation function. The first MSRD module takes the mask from the MPAM module and the feature map from the fifth layer as its input feature. The subsequent two MSRD modules are placed between the fourth and second layers of the decoder. In this configuration, the previous feature map obtained from upsampling is employed to generate the mask, which is then fed into the corresponding MSRD module. To achieve the upsampling process within the decoder,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 May 2024 doi:10.20944/preprints202405.1584.v1

deconvolution operations are employed. These operations enable the image resolution to be increased by a factor of 2 within each decoder block, effectively restoring finer details that may have been lost during the downsampling phase.

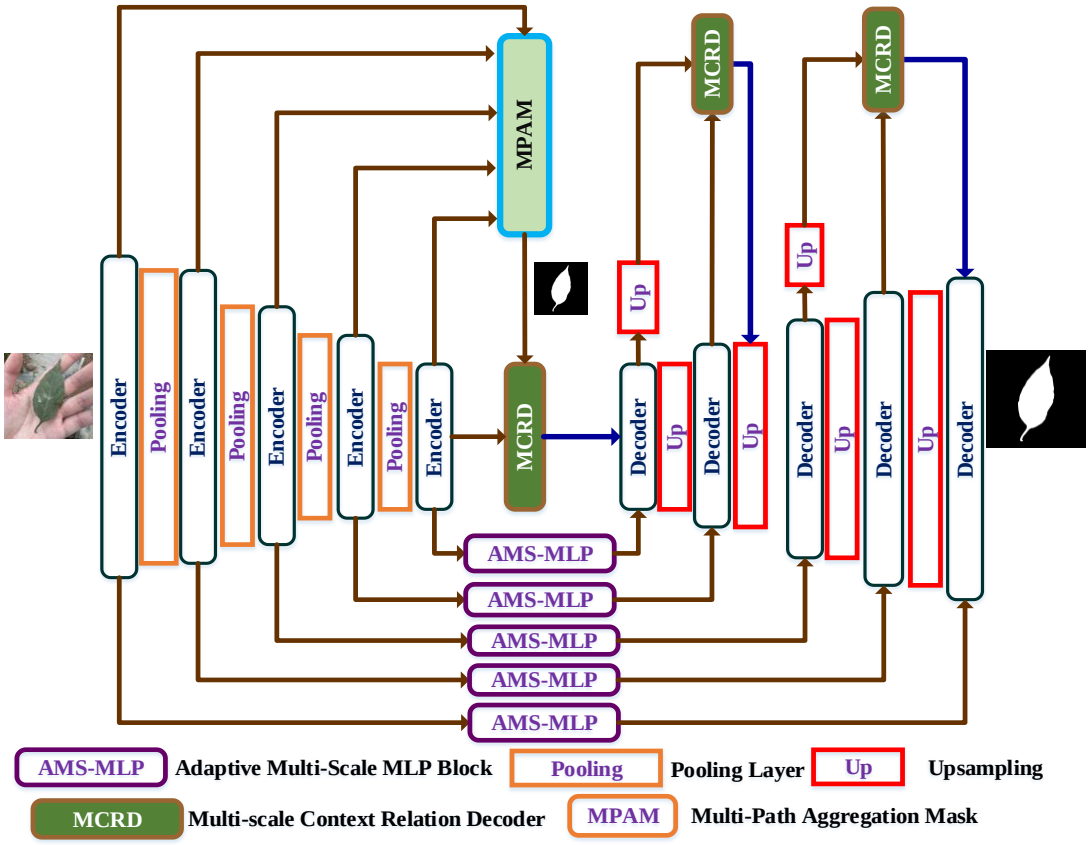

Figure 2. Overview of the AMS-MLP framework including the encoder network, the adaptive multi-scale MLP (AM-MLP) module, and the decoder network. The encoder network comprises five convolutional layers incorporating four downsampling operations and a multi-path aggregation mask (MPAM) module. The decoder network comprises five convolutional layers, incorporating four upsampling layers and three MSRD modules. The AM-MLP module is used for the skip connection layer.

3.2.2. Adaptive Multi-Scale MLP (AM-MLP) Module

The MLP module has demonstrated promising performance in the computer vision task, but it struggles with capturing spatial information and extracting global context due to its fully connected nature. To overcome these limitations, MAXIM [46] employs multi-scale MLP modules to extract global and local information. Inspired by MAXIM, we introduce an adaptive multi-scale MLP module that utilizes the self-attention mechanism to automatically extract multi-scale features and local information. As illustrated in Figure 3. the network initially splits the feature maps into two branches: the global multi-scale MLP (GMS-MLP) branch and the local multi-scale MLP (LMS-MLP) branch. The GMS-MLP branch focuses on extracting global features, while the LMS-MLP branch is dedicated to capturing local feature maps. Figure 4 illustrates the GMS-MLP and LMS-MLP modules.To effectively combine these features, we introduce an adaptive attention module that dynamically adjusts the weights of the global and local features based on their importance and relevance to the task. By incorporating the adaptive multi-scale MLP module, the AM-MLP module enabled the extraction of both global and local information in an adaptive manner while preserving spatial information and capturing contextual cues from different scales.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 May 2024 doi:10.20944/preprints202405.1584.v1

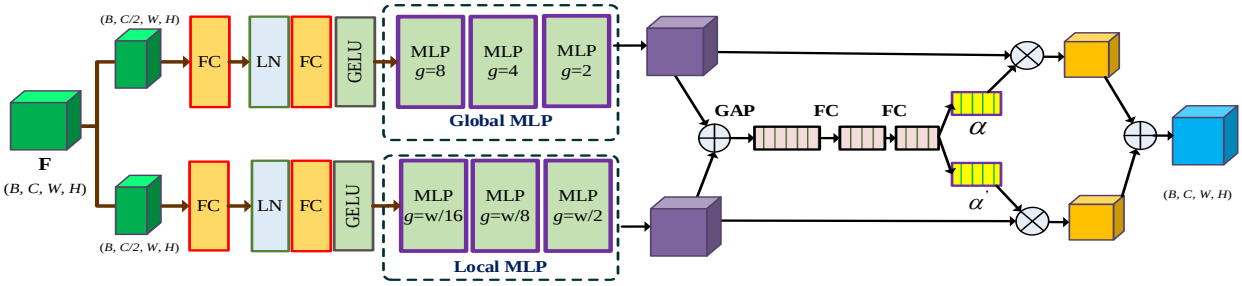

Figure 3. The network architecture of the AM-MLP module. The input feature map F is split into the global multi-scale MLP (GMS-MLP) branch FG and the local multi-scale MLP (LMS-MLP) branch FL. After each branch with multiple Cascade MLP blocks, the resulting features are alternately multiplied to enhance information interaction and then added together. Then, multi-scale features and local information are automatically extracted using an adaptive attention mechanism.

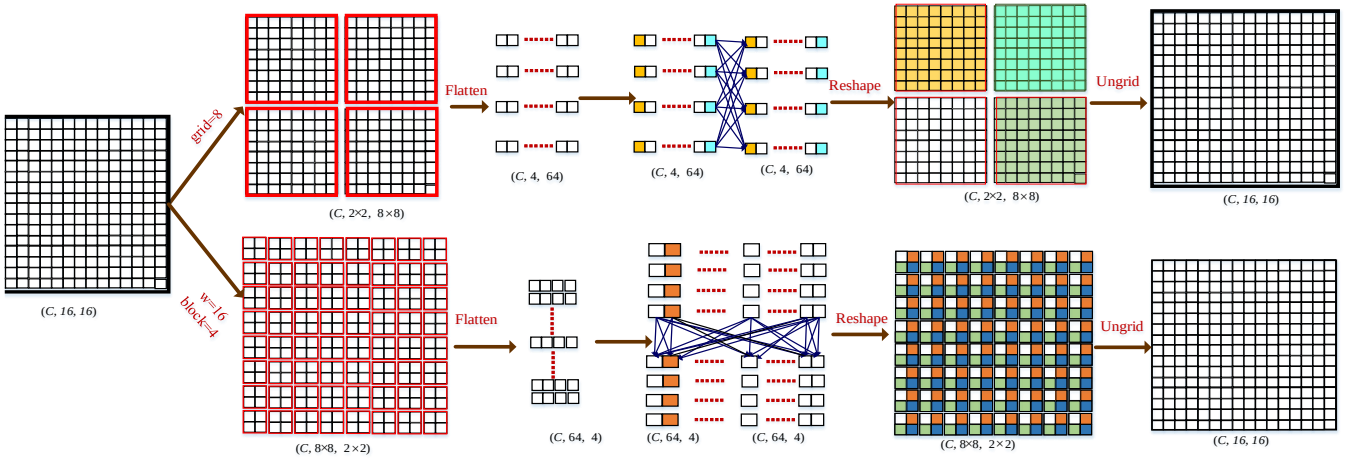

Figure 4. Illustration of the GMS-MLP and LMS-MLP modules. As an example, we used $\mathbf{F} \in \Re^{[B,\ C,\ H,\ W]}$ (W = 16; H = 16) as input, where B is the batch size, and C is the channel number. Input feature $\mathbf{F}$ will be processed by GMS-MLP and LMS-MLP branches. In the GMS-MLP branch, the feature map $\mathbf{F}^{\mathbf{G}}$ is initially divided into non-overlapping patches of size 2 × 2, resulting in a grid of size 8× 8. These patches are then flattened and fed into a fully connected (FC) layer along the first axis. Finally, the output is reshaped back and ungridded to restore the original size. In the LMS-MLP branch, the feature map $\mathbf{F}^{\mathbf{L}}$ is divided into non-overlapping patches of size 8 × 8, resulting in a blocking of size 2 × 2. These patches are flattened and processed through an FC layer along the second axis. Following that, the output is reshaped back and unblocked to regain the original size, resulting in the feature map $\mathbf{F}^{\mathbf{L}}_{\mathbf{mlp}}$.

Specially, the input features $\mathbf{F} \in \Re^{[B,\ C,\ H,\ W]}$ undergo an initial split into two branches based on the channel dimension, namely the GMS-MLP branch $\mathbf{F}^{\mathbf{G}} \in \mathbb{R}^{[B,\ C/2,\ H,\ W]}$ and the LMS-MLP branch $\mathbf{F}^{\mathbf{L}} \in \Re^{[B,\ C/2,\ H,\ W]}$, where B represents the batch size, C represents the channel number, and H and W represent the height and width of the image, respectively. In the GMS-MLP branch, the input features are first passed through a fully connected (FC) layer, followed by a layer normalization (LN) layer. The next step involves applying an additional fully connected (FC) layer and a GELU activation layer to generate the feature map $\mathbf{F}^{\mathbf{G}}_{\mathbf{fc}} \in \Re^{[B,\ C/2,\ H,\ W]}$. The generated feature map is then transformed into non-overlapping image patches, where each patch consists of a certain number of $g \times g$ grids. These patche features $\mathbf{F}^{\mathbf{G}}_{\mathbf{patch}} \in \Re^{B,\ C,\ g\times g,\ H_g \times W_g}$ are further processed through three consecutive multi-scale MLP modules, where $H_g = H/g$, $W_g = W/g$, and g is the kernel size. This process leads to the generation of novel feature maps $\mathbf{F}^{\mathbf{G}}_{\mathbf{mlp}}$, which can be denoted as follows:

$$\mathbf{F}^{\mathbf{G}}, \mathbf{F}^{\mathbf{L}} = split(\mathbf{F}) \qquad \mathbf{F} \in \Re^{B,\ C,\ H,\ W}, \mathbf{F}^{\mathbf{G}}, \mathbf{F}^{\mathbf{L}} \in \Re^{B,\ C/2,\ H,\ W} \tag{1}$$

$$\mathbf{F}_{\mathbf{fc}}^{\mathbf{G}} = fc(Ln(fc(Gelu(\mathbf{F}^{\mathbf{G}})))) \tag{2}$$

$$\mathbf{F}_{\mathbf{patch}}^{\mathbf{G}} = Reshape(\mathbf{F}_{\mathbf{fc}}^{\mathbf{G}}) \qquad F_{patch}^{G} \in \Re^{B,\ C,\ g\times g,\ H_g\times W_g} \tag{3}$$

$$\mathbf{F}_{\mathbf{mlp}}^{\mathbf{G}} = mlp_g(\mathbf{F}_{\mathbf{patch}}^{\mathbf{G}}) \qquad g \in [g_1,\ g_2,\ g_3] \tag{4}$$

where $split(\cdot)$ denoting dividing a multidimensional matrix or tensor into multiple sub-tensors along a channel dimension, and $fc(\cdot)$ denotes the full connection layer. $Ln(\cdot)$ denotes layer normalization layer, $Gelu(\cdot)$ denotes the GELU activation function, $Reshape(\cdot)$ denotes the operation of changing the shape or dimensions of two feature matrices. The GMS-MLP branch $mlp_g(\cdot)$ is three continuous MLP modules with the grid sizes of $g_1 \times g_1$, $g_2 \times g_2$, and $g_3 \times g_3$, respectively.

Similarly, in the LMS-MLP branch, the LMS-MLP feature $\mathbf{F}^{\mathbf{L}}$ passes through a FC layer, a layer normalization (LN) layer. Subsequently, it passes a FC layer and a GELU activation layer. The novel feature maps $\mathbf{F}_{\mathbf{Fc}}^{\mathbf{L}} \in \Re^{[B,\ C/2,\ H,\ W]}$ are projected into non-overlapping image patches and generate a new feature maps $\mathbf{F}_{\mathbf{block}}^{\mathbf{L}} \in \Re^{B,\ C,\ b\times b,\ H_b\times W_b}$, where H_b=H/b, W_b=W / b, b is the kernel size, and the size of each image patch is $b \times b$ grids. Then, the feature maps $\mathbf{F}_{\mathbf{fc}}^{\mathbf{L}}$ pass three continuous multi-scale MLP modules to obtain the spatial information, which is written as:

$$\mathbf{F}_{\mathbf{fc}}^{\mathbf{L}} = fc(Ln(fc(Gelu(\mathbf{F}^{\mathbf{L}})))) \tag{5}$$

$$\mathbf{F}_{\mathbf{block}}^{\mathbf{L}} = Reshape(\mathbf{F}_{\mathbf{fc}}^{\mathbf{L}}) \qquad F_{block}^{L} \in \Re^{[B,\ C,\ b\times b,\ H_b\times W_b]} \tag{6}$$

$$\mathbf{F}_{\mathbf{mlp}}^{\mathbf{L}} = mlp_b(\mathbf{F}_{\mathbf{block}}^{\mathbf{L}}) \qquad b \in [b_1,\ b_2,\ b_3] \tag{7}$$

where $mlp_b(\cdot)$ is three continuous MLP modules with the grid sizes of $b_1 \times b_1$, $b_2 \times b_2$, and $b_3 \times b_3$, respectively.

A self-attention module is employed to effectively fuse two features $\mathbf{F}_{\mathbf{mlp}}^{\mathbf{L}}$ and $\mathbf{F}_{\mathbf{mlp}}^{\mathbf{L}}$ obtained from the GMS-MLP and LMS-MLP branches, and it guides the segmented network to select more representative features from the channel dimension. Specially, two features $\mathbf{F}_{mlp}^{G}$ and $\mathbf{F}_{mlp}^{L}$ are fused, and followed by the global average pooling (GAP) operation to compress the channel dimension, which can be represented as follows:

$$\mathbf{F}_{\mathrm{H}}^{\mathrm{G}} = GAP(\mathbf{F}_{mlp}^{\mathrm{G}} \oplus \mathbf{F}_{mlp}^{\mathrm{L}}) \tag{8}$$

where $\mathbf{F}_{\mathbf{H}}^{\mathbf{G}}$ is the output features of the GAP layer. Then, the features $\mathbf{F}_{\mathbf{H}}^{\mathbf{G}}$ are input into a FC layer, followed by a batch normalization layer, and a softmax function. The probability feature maps $\mathbf{F}_{\mathbf{H}}^{\mathbf{FC}}$ can be expressed as:

$$\mathbf{F}_{\mathrm{H}}^{\mathrm{FC}} = \sigma\left(BN(fc(\mathbf{F}_{\mathrm{H}}^{\mathrm{G}}))\right) \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $BN(\cdot)$ is a batch normalization layer. Then, we perform another FC layer on the features $\mathbf{F}_{\mathbf{H}}^{\mathbf{FC}}$ followed by the softmax activation function, and the channel attention map $\boldsymbol{\alpha}$ is written as:

 doi:10.20944/preprints202405.1584.v1

$$\boldsymbol{\alpha} = \sigma_{sf}\left(fc(\mathbf{F}_{\mathrm{H}}^{\mathrm{FC}})\right) \tag{10}$$

where $\sigma_{sf}(\cdot)$ denotes the softmax activation layer. We regard the channel attention map $\boldsymbol{\alpha} \in [0, 1]$ as the weight of the features, where $\alpha \in \Re^{C\times 1\times 1}$. The channel attention map $\boldsymbol{\alpha}^{'} \in [0, 1]$ is from the value α, and it satisfies $\boldsymbol{\alpha}^{'} = 1 - \boldsymbol{\alpha}$. An important observation is that the channel attention maps α and $\alpha^{'}$ enable the adaptive adjustment of weights for the two channel attention feature maps. It also demonstrates that the two feature maps are capable of extracting feature representations from different receptive fields. By flexibly adjusting the adaptive weights of two features $\mathbf{F}_{\mathbf{Fc}}^{\mathbf{G}}$ and $\mathbf{F}_{\mathbf{Fc}}^{\mathbf{L}}$, the feature maps can be expressed as follows:

$$\mathbf{F}_{\mathbf{G}}^{'} = \boldsymbol{\alpha} \times \mathbf{F}_{\mathbf{FC}}^{\mathbf{G}} \tag{11}$$

$$\mathbf{F}_{\mathbf{L}}^{'} = \boldsymbol{\alpha}^{'} \times \mathbf{F}_{\mathbf{FC}}^{\mathbf{L}} \tag{12}$$

$$\mathbf{F}^{\mathbf{out}} = \mathbf{F}_{\mathbf{G}}^{'} \odot \mathbf{F}_{\mathbf{L}}^{'} \tag{13}$$

where $\odot$ denotes the concatenation operator, $\mathbf{F}_{\mathrm{G}}^{'}, \mathbf{F}_{\mathrm{L}}^{'} \in \Re^{[\mathrm{B},\ \mathrm{C}/2,\ \mathrm{H},\ \mathrm{W}]}$ are two output features from the adaptive dot-product features, respectively.

Notably, the grid size g and the block size b satisfy a specific relationship. As exemplified in Figure 3, when reducing the patch size in the GMS-MLP block, the block size in the LMS-MLP branch increases accordingly. For instance, when considering an image size of 32, the grid sizes in the GMS-MLP branch are set to 8, 4, and 2, while the corresponding grid sizes in the LMS-MLP branch are 4, 8, and 16, respectively. This arrangement results in a larger number of patches within the global MLP, enabling the capture of spatial information among the patches. Conversely, in the LMS-MLP branch, a larger number of pixels in each block allows for the retention of local spatial information between pixels. Consequently, by fusing the GMS-MLP and LMS-MLP blocks, a comprehensive feature map can be generated, encompassing both global and local information in a progressively richer manner.

3.2.3. Multi-Scale Context Relation Decoder (MCRD) Module

The accurate extraction of boundaries between foreground and background regions relies on the presence of both local and contextual information. To address this, the Mask refinement network [49] leverages contextual relationships to improve the pixel boundaries in these regions. In line with this, we propose an MCRD module to enhance the target boundary features and contextual information. As shown in Figure 5, our approach involves initial upsampling of the high feature maps $\mathbf{F}^{\mathrm{i+1}}$ through non-linear interpolation with a rate of 2, followed by a sigmoid activation function. The novel feature maps are then fed into a 1×1 convolutional block, which generates an output with a single channel. The process is formulated as follows:

$$\mathbf{F}^{\mathrm{up}} = up(\mathbf{F}^{\mathrm{i+1}}) \tag{14}$$

$$\mathbf{F}^{\mathrm{mask}} = Conv_{1\times 1}\left(\sigma\left(\mathbf{F}^{\mathrm{up}}\right)\right) \tag{15}$$

where $up(\cdot)$ denotes the upsampling operator, $Conv_{1\times 1}(\cdot)$ is a 1×1 convolutional operation.

Then, the mask maps $\mathbf{F}^{\mathrm{mask}}$ is used to assign different weights of the foreground and background feature maps, which are written as:

$$\mathbf{F}^{\mathrm{fg}} = Conv_{3\times 3}(\mathbf{F}^{\mathrm{i}} \otimes \mathbf{F}^{\mathrm{mask}}) \tag{16}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 May 2024 doi:10.20944/preprints202405.1584.v1

$$\mathbf{F}^{bg} = Conv_{3\times3}\left(\mathbf{F}^{i} \otimes \left(1-\mathbf{F}^{mask}\right)\right) \tag{17}$$

where $\otimes$ denotes the dot product, $Conv_{3\times3}(\cdot)$ is a 3×3 convolutional block.

Finally, we concatenate two feature maps $\mathbf{F}^{fg}$ and $\mathbf{F}^{bg}$ on the channel dimension, and it then perform a 3×3 convolutional layer, which is written as:

$$\mathbf{F}^{bg} = Conv_{3\times3}\left(\mathbf{F}^{fg} \odot \mathbf{F}^{bg}\right) \tag{18}$$

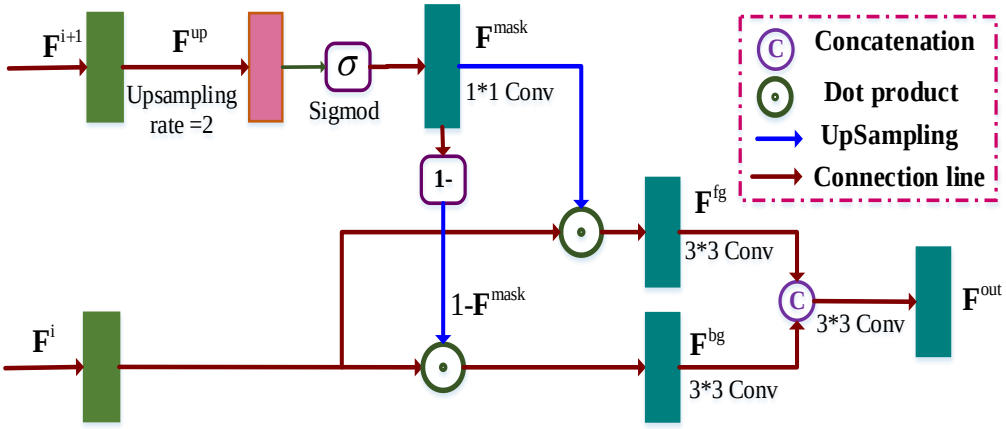

Figure 5. Illustration of the multi-scale context relation decoder (MCRD) module. Two feature maps $\mathbf{F}^{i}$ and $\mathbf{F}^{i+1}$ are input into the MCRD module, the high features is first performed on the upsampling operation. The generated feature maps $\mathbf{F}^{up}$ pass through the sigmod activation function and a 1×1 convolutional operation, which generates the mask maps $\mathbf{F}^{mask}$ representing the foreground and background regions. .

3.2.4. Multi-Path Aggregation Mask (MPAM) Module

The multi-scale nature of features in deep neural networks offers different levels of information, with deeper layers capturing coarser details and shallower layers preserving finer details. To leverage the benefits of each layer, we introduce an MPAM module to enhance the extraction of accurate boundary information and facilitate the generation of masks. As shown in Figure 6, for the feature map $\mathbf{F}_i$ from the fifth layer to the second layer in the encoder, each feature map is subjected to a 1×1 convolutional operation to decrease the channel dimensions. The resulting feature maps have the same channel number as the first layer in the encoder. Additionally, we employ an upsampling operation with a rate of 2 on these feature maps. This procedure can be expressed as follows:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 May 2024 doi:10.20944/preprints202405.1584.v1

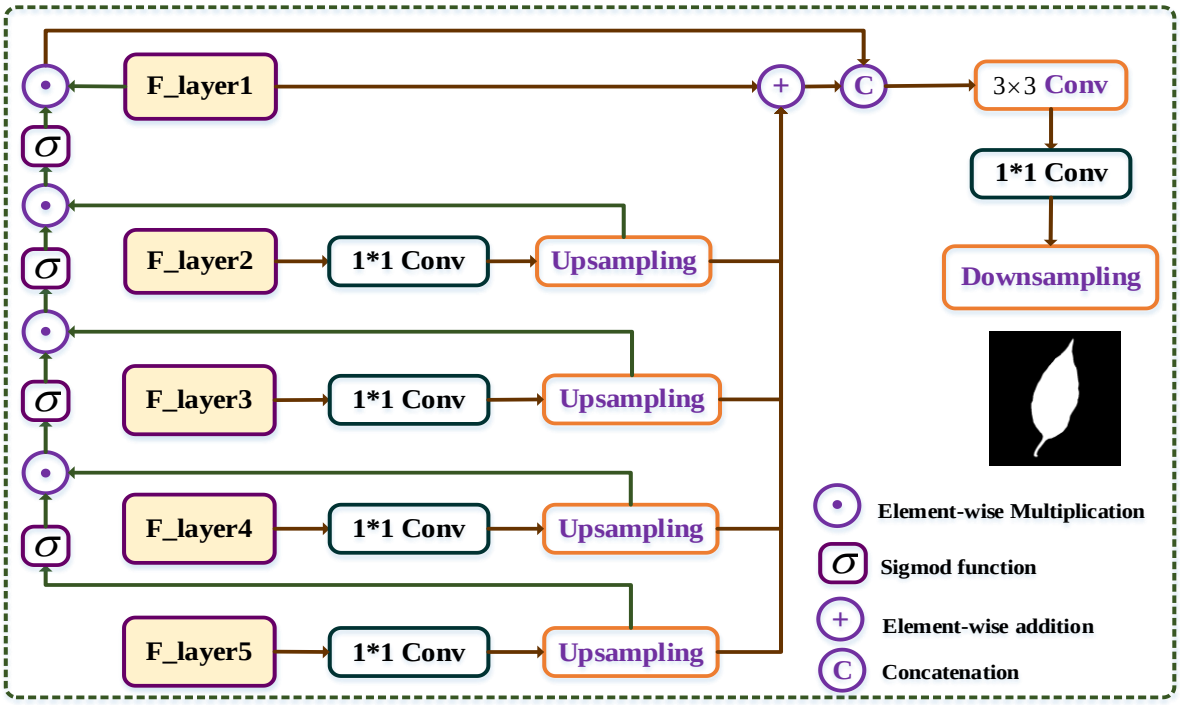

Figure 6. Illustration of the multi-path mask decoder module. From the fifth to second layers, the feature maps $\mathbf{F}_i$ in the encoder are first passing a 1×1 convolutional operation to suppress the channel number, and the generated channel number of the output features is the same to that of the first layer in the encoder.

$$\mathbf{F}_{i-1}^{up} = up\left(Conv_{1\times1}\left(\mathbf{F}_i\right)\right) \qquad i = 2,\cdots5 \tag{19}$$

Subsequently, the generated feature maps are further processed by the Sigmoid activation function. We then concatenate the generated feature $\mathbf{F}_{i-1}^{up}$ and the previous feature maps $\mathbf{F}_{i-1}$, and the final feature maps are written as:

$$\mathbf{F}_{i-1}^{out} = \sigma\left(\mathbf{F}_{i-1}^{up}\right)\odot\mathbf{F}_{i-1} \qquad i = 2,\cdots5 \tag{20}$$

To incorporate information from various scales, we utilize an element-wise addition operation between four upsampling feature maps and the feature maps $\mathbf{F}_{i-1}^{up}$ obtained from the first layer in the encoder. This operation produces multi-scale fusion feature maps (MSFF), which can be denoted as:

$$\mathbf{F}^{cat} = \{\mathbf{F}_i^{up}\odot\}_{i=1}^{4}\mathbf{F}_1 \tag{21}$$

In the final step, we concatenate the MSFF maps $\mathbf{F}^{cat}$ with the feature maps $\mathbf{F}_i^{out}$ obtained from the first layer. The concatenated feature maps $\mathbf{F}_1^{en}$ are then fed into a convolutional block with 3×3 filters. To generate the mask for the foreground and background regions, we apply a 1×1 convolutional operation with a single output channel, followed by a downsampling operation. Mathematically, this can be represented as:

$$\mathbf{F}_1^{en} = \mathbf{F}^{cat}\odot\mathbf{F}_1^{out} \tag{22}$$

$$\mathbf{F}_{i-1}^{up} = dn\left(Conv_{3\times3}\left(\mathbf{F}_i\right)\right) \qquad i = 2,\cdots5 \tag{23}$$

where $dn(\cdot)$ denotes the downsampling operator.

 doi:10.20944/preprints202405.1584.v1

3.2.5. Training Loss

The proposed AMS-MLP network involves two loss functions to optimize the predicted result and the ground truth (GT), including the binary cross entropy (BCE) L_b and the Dice L_d. The two loss functions are defined as:

$$L_b(f,g) = -\sum_{i=1}^{N}[g_x \log(f_x) + (1-g_x)\log(1-f_x)] \tag{24}$$

$$L_d(f,g) = 1 - \frac{2\sum_{i=1}^{N} f_x \cdot g_x}{\sum_{i=1}^{N} f_x + \sum_{i=1}^{N} g_x} \tag{25}$$

where f denotes the input predicted result, and g denotes the corresponding ground truth label.Therefore, our final loss L_{loss} can be expressed as:

$$L_{loss}(f,g) = \alpha L_{bce}(f,g) + L_d(f,g) \tag{26}$$

3.2.6. Performance Evaluation

To rigorously evaluate the performance of the proposed method and other compared methods, six metrics are employed as evaluation criteria: accuracy, recall, precision, specificity, F1-score, and intersection over union (IoU). The first five metrics are widely used measures, and they are defined as:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{28}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{29}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{30}$$

where TP (true positive) represents the number of pixels that are correctly predicted as foreground, TN (true negative) indicates the number of pixels that are correctly predicted as background, FP (false positive) refers to the number of pixels that are predicted as foreground but actually belong to the background according to the ground truth. On the other hand, FN (false negative) represents the number of pixels that are predicted as background but actually belong to the foreground according to the ground truth.

The F1-score by combining two indices PR and PP is defined as:

$$F1 = \frac{2 \times PR \cdot PP}{PR + PP} \tag{31}$$

The IoU is a metric commonly used to evaluate the accuracy of boundary predictions. It measures the overlap between the predicted border and the real border by calculating the ratio of their intersection to their union. Mathematically, the IoU is defined as follows:

 doi:10.20944/preprints202405.1584.v1

$$IoU = \frac{TP}{TP + FP + FN} \tag{32}$$

4. Experiments

4.1. Experimental Settings

The proposed AMS-MLP network was implemented on a specific hardware configuration consisting of a 12th Gen Intel(R) Core(TM) i7-12700K 3.60 GHz processor and an NVIDIA GeForce RTX 3090 40 GB GPU with 32 GB of RAM. The operating system employed was Windows 11, and the Conda environment was utilized to ensure a consistent software environment for the execution. Regarding the parameter settings for the AMS-MLP network, the following values were employed: a batch size of 2 and a total of 60 epochs. Throughout the training process, the network optimization was performed using stochastic gradient descent (SGD) as the optimization algorithm, with an initial learning rate of 0.001. These parameter choices were made to facilitate effective training and convergence of the AMS-MLP network.

4.2. Comparison with the SOTA Models

In order to assess the segmentation performance of the AMS-MLP model, we conducted a comparative evaluation against state-of-the-art (SOTA) models on three distinct leaf datasets: EBD, BSD and MLD. The compared models included FCN-VGG16, U-Net [25], attention U-Net (AttU-Net) [26], UNet++ [27], UNeXt [47], and CM-MLP model [48]. To ensure a fair and comprehensive comparison, all models were trained, validated, and tested on the same three datasets. By maintaining consistency across the training, validation, and test datasets, we aim to eliminate any potential bias or variation that may affect the results and evaluate the segmentation performance exclusively on the test dataset.

Table 2 presents the segmentation results of the proposed model in comparison to seven segmentation models on the EBD dataset. The evaluation is based on six indices: accuracy, recall, precision, mean IoU (mIoU), and F1-score. Among five evaluation indices, namely accuracy, recall, precision mIoU, and F1 scores, the proposed model achieves the highest scores. Notably, the mIoU and F1-score of our model are 97.39% and 98.29%, respectively, surpassing the performance of the FCN model by 0.35% and 0.29%. In comparison to U-Net, our proposed model demonstrates significant improvements, with an increase of 9.92% in mIoU, 0.18% in F1-score, and 0.29% in recall. Furthermore, when compared to other semantic segmentation models, our proposed model achieves the highest scores in accuracy, recall, precision, mIoU, and F1-score. To provide a comprehensive assessment, qualitative examples of segmentation results are presented. Figure 7 illustrates the capability of all models to accurately locate the object region, while our proposed model stands out in accurately delineating the boundaries of pepper leaves in the given test images. These results further substantiate the effectiveness and superiority of our proposed model in the task of leaf segmentation.

Table 2. The results of segmenting the EBD dataset using seven different models.

Model		Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	mIoU (%)	F1-score (%)
FCN	**FCN-16s**	99.45	97.33	99.79	98.67	97.04	98.00
	U-Net	99.53	97.31	**99.85**	98.92	87.47	98.11
UNet-based	**AttU-Net**	99.26	96.29	99.74	98.38	96.06	97.33
	UNet++	99.43	97.04	99.82	98.87	96.95	97.94
	UNeXt	99.31	96.31	99.79	98.67	96.38	97.48
MLP-based	**CM-MLP**	99.44	97.41	99.77	98.54	96.96	97.97
	Ours	**99.53**	**97.61**	99.84	**98.97**	**97.39**	**98.29**

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 May 2024 doi:10.20944/preprints202405.1584.v1

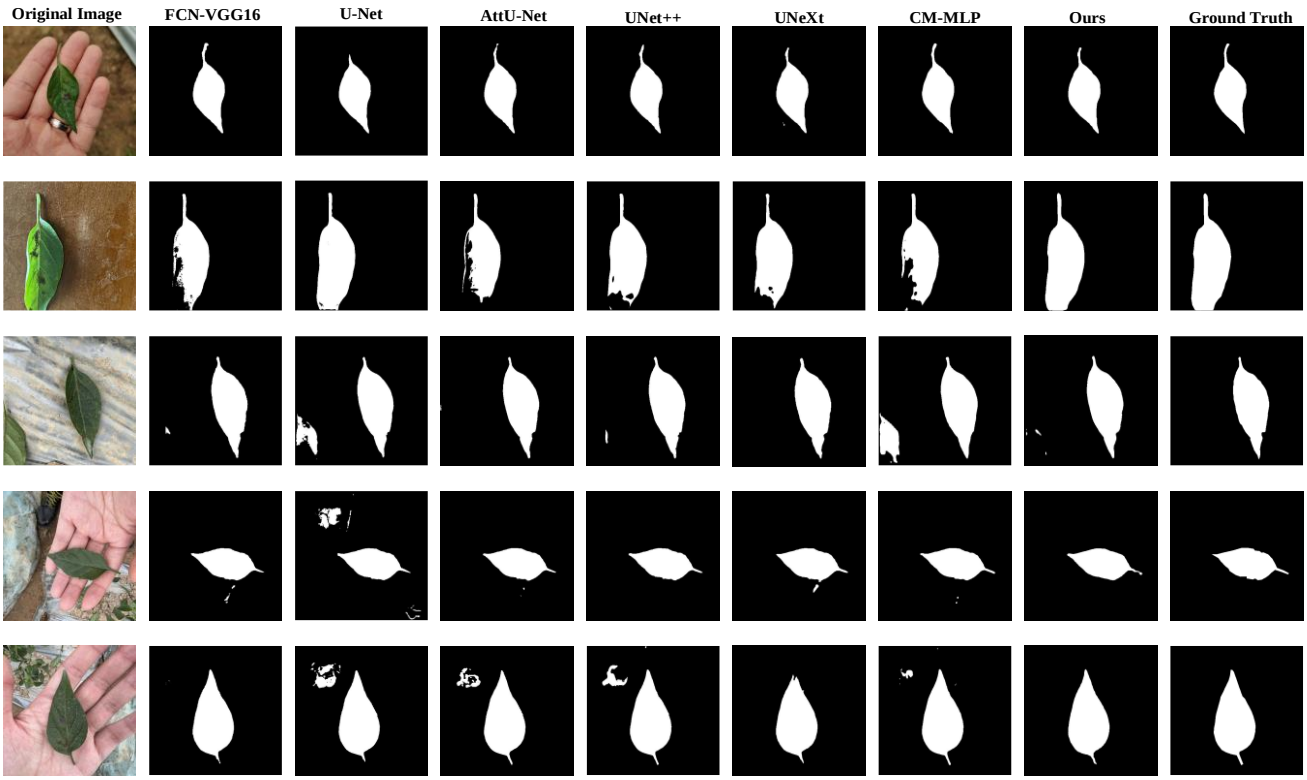

Figure 7. Qualitative comparison of the proposed model compared with six models on the EBD dataset, and five examples of the predicted results are shown. From the 1st column to 9th column: the original image, the predicted results corresponding to FCN-VGG16, U-Net, AttUNet, UNet++, UNeXt, CM-MLP, our model, and the ground truth, respectively.

To evaluate the training performance of the AMS-MLP network, we compared it with FCN-based, UNet-based, and MLP-based models on the BSD and MLD datasets. The FCN-based models included FCN-16s, the UNet-based models included U-Net, AttU-Net, and UNet++, and the MLP-based models included UNeXt and CM-MLP. From Table 3 and Table 4, our method achieved the best segmentation performance across five evaluation metrics. This improvement can be attributed to the incorporation of GMS-MLP and LMS-MLP in the AM-MLP network, which can extract both global and local information from the target regions, thereby enhancing its segmentation capability. The FCN-16s model exhibits superior performance compared to U-Net, which can be attributed to its utilization of VGG16 as a pretrained model for the feature extraction. By incorporating the pretrained VGG16 model, the FCN-16s model can extract more comprehensive and informative features, leading to a richer feature representation within the encoder. The CM-MLP model replaces the attention mechanism with MLP and achieves better segmentation results than the U-Net model by considering the relationships between pixels. From Table 3 and Table 4, it can be observed that our model achieves the highest mIoU and F1 scores, indicating its superior segmentation performance among the evaluated models.

Table 3. The results of segmenting the BSD dataset using seven different models.

Model		Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	mIoU (%)	F1-score (%)
FCN	**FCN-16s**	99.69	97.17	99.87	98.11	96.62	97.64
UNet-based	**U-Net**	98.83	93.85	99.18	88.95	96.14	91.33
	AttU-Net	99.62	98.05	99.73	96.26	95.97	97.14
	UNet++	99.37	**98.08**	99.46	92.68	95.75	95.31
MLP-based	**UNeXt**	99.37	97.70	99.49	93.06	94.66	95.33
	CM-MLP	99.66	96.95	99.85	97.83	95.68	97.39

 doi:10.20944/preprints202405.1584.v1

	Ours	**99.72**	97.47	**99.88**	**98.26**	**96.91**	**97.86**

Table 4. The results of segmenting the MLD dataset using seven different models.

Model		Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	mIoU (%)	F1-score (%)
FCN	**FCN-16s**	99.61	96.93	99.89	98.94	97.10	97.92
	U-Net	99.46	95.40	99.88	98.80	96.19	97.07
UNet-based	**AttU-Net**	99.57	96.43	99.90	99.03	97.05	97.71
	UNet++	98.87	89.58	99.83	98.25	92.15	93.71
	UNeXt	99.20	92.84	99.86	98.56	94.24	95.61
MLP-based	**CM-MLP**	99.71	**98.02**	99.88	98.85	97.32	98.44
	Ours	**99.72**	97.79	**99.92**	**99.24**	**97.91**	**98.51**

Figures 8 and 9 demonstrate that the predicted images segmented by U-Net and UNet++ exhibit certain false positive regions, where the lesion areas are not effectively differentiated from the background. However, AttUNet can obtain better segmentation results compared to U-Net. Conversely, the lesion images segmented by the AM-MLP model closely resemble the ground truth, exhibiting precise extraction of pepper leaf boundaries and a significantly reduced false positive region in comparison to other networks. This superior performance can be attributed to the incorporation of the MCRD module and the utilization of the GMS-MLP and LMS-MLP auxiliary streams, which facilitate cascaded contraction and expansion within the network.

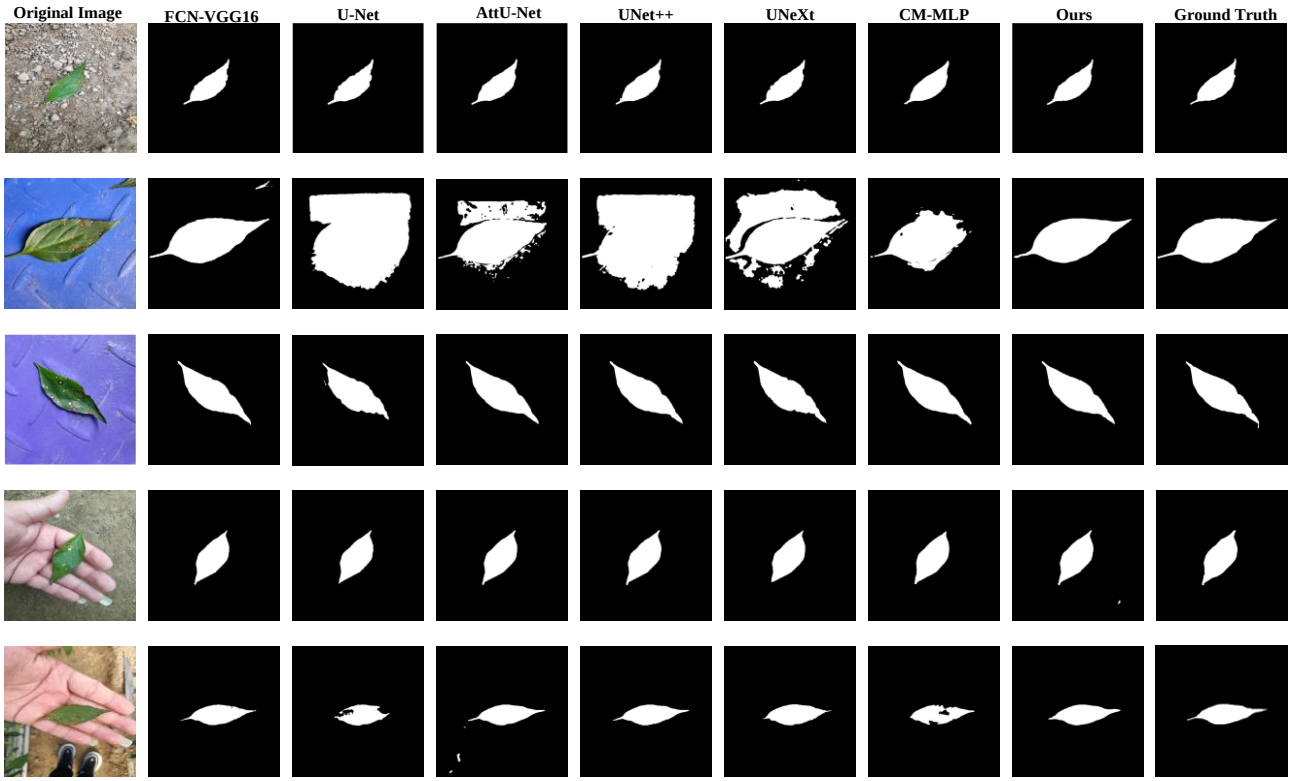

Figure 8. Qualitative comparison of the proposed model compared with six models on the BSD dataset, and five examples of the predicted results are shown. From the 1st column to 9th column: the original image, the predicted results corresponding to FCN-VGG16, U-Net, AttUNet, UNet++, UNeXt, CM-MLP, our model, and the ground truth, respectively.

 doi:10.20944/preprints202405.1584.v1

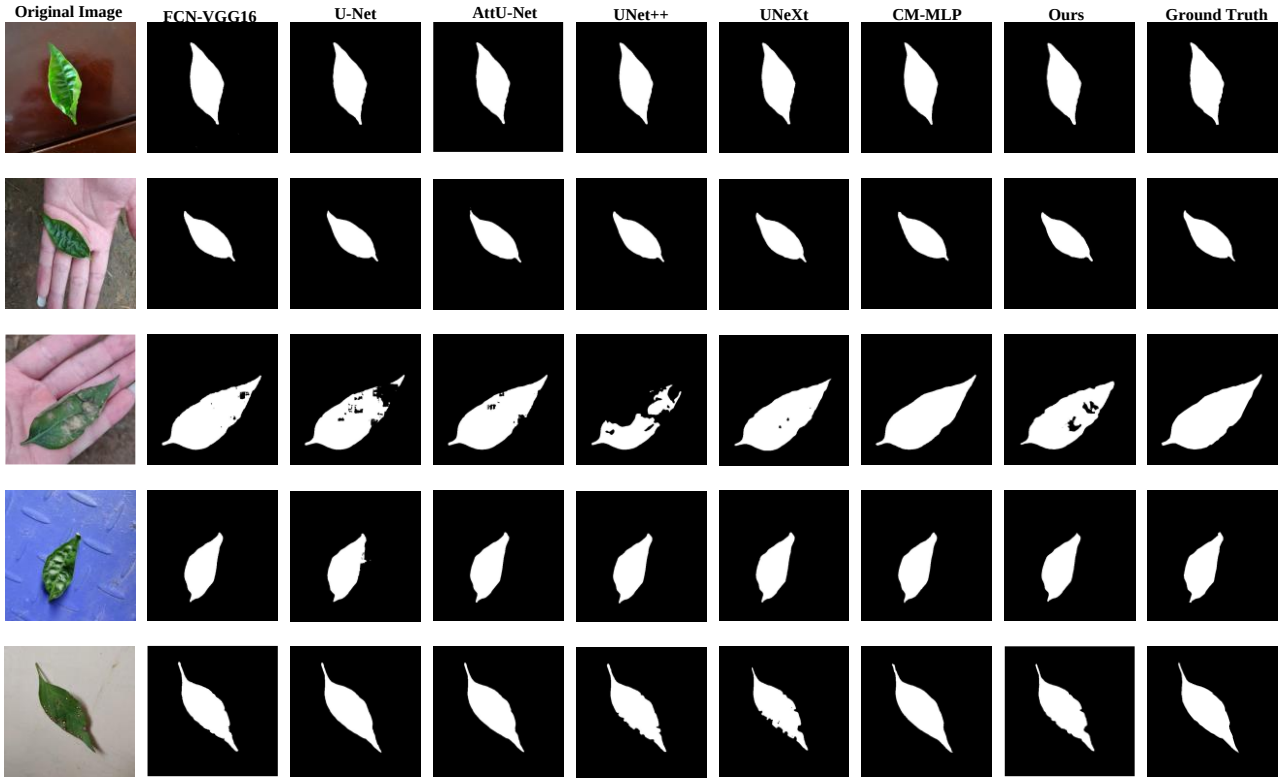

Figure 9. Qualitative comparison of the proposed model compared with six models on the MLD dataset, and five examples of the predicted results are shown. From the 1st column to the 9th column: the original image, the predicted results corresponding to FCN-VGG16, U-Net, attention U-Net (AttUNet), UNet++, UNeXt, CM-MLP, our model, and the ground truth, respectively.

4.3. Ablation Study

In this section, we conducted an ablation experiment to evaluate the effectiveness of individual modules on the MLD dataset. Our baseline model was constructed based on the U-Net architecture, with a reduced number of channels compared to the original U-Net. To evaluate the segmentation performance of each module, a step-wise approach was employed. Initially, we introduced the BAM-MLP (BU-Net+AM-MLP) network, which incorporated an AM-MLP module to capture global context features. The AM-MLP module enhanced attention towards informative regions, thereby improving the overall segmentation performance. Subsequently, we integrated the MPAM module into the fifth layer of the encoder, resulting in the BMAM-MLP (BU-Net+ MPAM+AM-MLP) network. The MPAM module played a crucial role in generating masks, thereby facilitating the refinement of the segmentation process. Finally, our comprehensive model (AMS-MLP) was formulated by progressively incorporating the MCRD module into the BMAM-MLP network. The MCRD module facilitated the extraction of boundary information, contributing to further improvements in segmentation accuracy. Through this incremental approach, we aimed to systematically evaluate the impact of each module on the overall segmentation performance. By analyzing the results, we can gain insights into the individual contributions of these modules and assess their effectiveness in enhancing the performance of the model.

 doi:10.20944/preprints202405.1584.v1

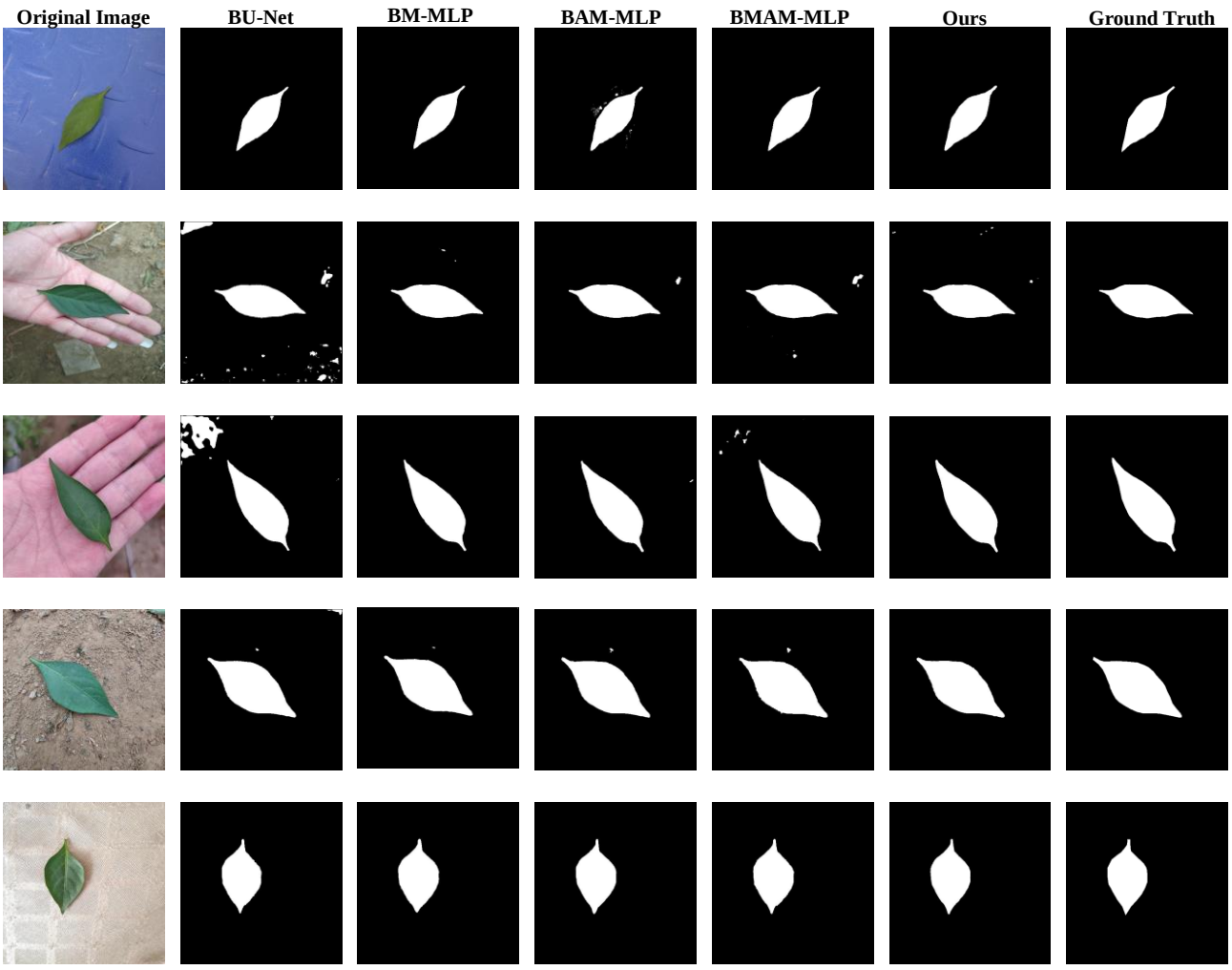

Figure 10. Qualitative comparison for the ablation study on the MLD dataset, and five predicted results are shown. From the 1st column to 9th column: the original image, the predicted results corresponding to BU-Net, BM-MLP, BAM-MLP, BMAM-MLP, our model, and the ground truth, respectively.

As presented in Table 5, our investigation begins with the BM-MLP model, which demonstrates superior performance compared to BU-Net. Notably, the BM-MLP model achieves accuracy, recall, precision, specificity, mIoU, and F1-score, with values of 99.41%, 97.38%, 95.02%, 99.58%, 97.00%, and 96.18%, respectively. Subsequently, we further enhance the BU-Net model by incorporating the AM-MLP module, leading to the development of the BAM-MLP model. Comparative analysis of the evaluation metrics demonstrates substantial improvements achieved by the BAM-MLP model over the BU-Net model. The BAM-MLP model achieves accuracy, recall, precision, mIoU, and F1-score of 98.37%, 96.70%, 84.20%, 98.51%, and 90.02%, respectively, surpassing the performance of BU-Net by increments of 0.72%, 2.78%, 0.55%, 0.32%, and 4.14%, respectively. Furthermore, we explore the synergistic effects obtained by integrating the AM-MLP module and MPAM module into the BU-Net model, resulting in the development of the BMAM-MLP model. The findings of our study reveal that the BMAM-MLP model significantly enhances the segmentation performance across multiple metrics, including accuracy, recall, precision, specificity, and F1-score. Lastly, we construct an AMS-MLP network by incorporating the MCRD module into the BMAM-MLP network. The MCRD module effectively utilizes multiscale features to preserve the boundary features. Consequently, the combination of these three modules, i.e., AM-MLP, MPAM, and MCRD, results in the best overall performance. From a comprehensive analysis of the combined metrics, it is evident that each module exhibits notable advantages when applied to the MLD dataset.

 doi:10.20944/preprints202405.1584.v1

Table 5. The Compared results for the ablation experiment of pepper leaf segmentation.

Model	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	mIoU (%)	F1-score (%)
Baseline (BU-Net)	97.65	93.92	79.11	97.96	97.74	85.88
BAM-MLP (BU-Net+AM-MLP)	98.37	96.70	84.20	98.51	98.06	90.02
BMAM-MLP (BU-Net+MPAM+AM-MLP)	99.04	96.78	91.36	99.25	96.88	93.87
Ours (BU-Net+MPAM+AM-MLP+MCRD)	**99.63**	**97.98**	**97.20**	**99.77**	**98.28**	**97.59**

5. Conclusion

The robust performance of current plant disease recognition models is poor due to the diversity of plant leaf image backgrounds. Therefore, accurate extraction of plant leaves from the background is highly desirable for plant disease recognition. In this paper, we propose a lightweight and accurate leaf segmentation model for extracting pepper leaves from diverse backgrounds. Specifically, we design an adaptive multi-scale MLP network by combining the MPAM module and the MCRD module for pepper leaf segmentation. It consists of an encoder network, an AM-MLP module, and a decoder network. Within the encoder network, the MPAM module is employed to aggregate features from five layers and generate a single-channel mask, enhancing the accuracy of pepper leaf boundary extraction. In the AM-MLP module, the GMS-MLP branch extracts global features while the LMS-MLP branch focuses on capturing local feature maps. Additionally, we employ an adaptive attention module to dynamically extract the features of the global and local branches. The decoder network incorporates the MCRD module into the convolutional layer, which enhances the ability of boundary extraction. To validate the generalizability of our proposed approach, extensive experiments are conducted on three pepper leaf datasets, and the results reveal mIoU scores of 97.39%, 96.91%, and 97.91%, as well as F1-score of 98.29%, 97.86%, and 98.51%, respectively. Meanwhile, the ablation experiments are conducted by gradually integrating three modules, namely AM-MLP, MPAM, and MCRD, into the baseline model. The results in Table 5 demonstrate significant improvements in segmentation performance across six evaluation metrics.

Although our proposed AMS-MLP network is able to segment pepper leaves effectively, the method is based on full supervision and requires a large number of training samples with labeling. In future work, we plan to explore a weakly supervised or self-supervised segmentation method for pepper leaf segmentation. On the other hand, we investigate fine-tuning methods on existing deep learning-based models to enhance the generalisation ability of the models and provide more effective and feasible solutions for pepper leaf images in different scenarios

Supplementary Materials: The code will be available at: https://github.com/fangchj2002/AMS-MLP.

Author Contributions: All authors contributed to the article and JF: conceptualization, methodology, experiment, and writing. JY: experiment and writing. HL: supervision and writing- review & editing.YF: methodology and approved the submitted version.

Funding: The research described in this paper was funded by the National Natural Science Foundation of China (No. 61966001, No.62206195, No. 61866001, No. 62163004, No. 61963002, and No. 62206195), the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (No. LZY23F050001), Natural Science Foundation of Jiangxi Province (No. 20202BABL214032 and No. 20202BABL203035), Science and Technology Plan Project of Taizhou City (No. 22ywa58 and No. 22nya18), Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology (No.JELRGBDT202201), the Engineering Research Center of Development and Management for Low to Ultra-Low Permeability Oil & Gas Reservoirs in West China(No. KFJJ-XB- 2020-1), and the Open Fund of Key Laboratory of Exploration Technologies for Oil and Gas Resources (No. K2021-02).

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

 doi:10.20944/preprints202405.1584.v1

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Bhavini, J., Sheshang, D.. A Survey on apple fruit diseases detection and classifcation. Int. J. Comput. Appl., 2015, 130 (13): 25-32.
2. Cruz, A., Ampatzidis, Y., Pierro, R., Materazzi, A., Panattoni, A., De Bellis, L., Luvisi, A.. Detection of grapevine yellows symptoms in Vitis vinifera L. with artificial intelligence. Comput. Electron. Agric., 2019, 157: 63-76.
3. Zhu, S., Ma, W., Lu, J., Ren, B., Wang, C., Wang, J.. A novel approach for apple leaf disease image segmentation in complex scenes based on two-stage DeepLabv3+ with adaptive loss. Comput. Electron. Agric., 2023, 204, 107539.
4. Pan, D., He, M., Kong, F.. Risk attitude, risk perception, and farmers' pesticide application behavior in China: A moderation and mediation model. J. Clean. Prod., 2020, 276, 124241.
5. Mu, H., Wang, K., Yang, X., Xu, W., Liu, X., Ritsema, C.J., Geissen, V.. Pesticide usage practices and the exposure risk to pollinators: A case study in the north China plain. Ecotoxicol. Environ. Saf., 2022, 241, 113713.
6. Luo, Y., Sun, J., Shen, J., Wu, X., Wang, L., Zhu, W.. Apple leaf disease recognition and sub-class categorization based on improved multi-scale feature fusion network. IEEE Access, 2021, 9: 95517-95527.
7. Liu, B., Zhang, Y., He, D., Li, Y.. Identification of apple leaf diseases based on deep convolutional neural networks. Symmetry, 2017, 10 (1), 11.
8. Beikmohammadi, A., Faez, K., Motallebi, A.. SWP-LeafNET: A novel multistage approach for plant leaf identification based on deep CNN, Expert Syst. Appl., 2022, 202, 117470.
9. Naik, B.N., Malmathanraj, R., Palanisamy, P.. Detection and classification of chilli leaf disease using a squeeze-and-excitation-based CNN model. Eco. Inform., 2022., 69, 101663.
10. Shafik W., Tufail A., Liyanage C.S., Apong R.A.A.H.M.. Using a novel convolutional neural network for plant pests detection and disease classification. J. Sci. Food Agric., 2023, 103(12): 5849-5861.
11. Pal, A., Kumar, V.. AgriDet: Plant leaf disease severity classification using agriculture detection framework. Eng. Appl. Artif. Intel., 2023, 119, 105754.
12. Alshammari, H. H., Taloba, A. I., Shahin, O.R.. Identification of olive leaf disease through optimized deep learning approach. Alex. Eng. J., 2023, 72: 213-224.
13. Yu, M., Ma, X.D., Guan, H.O., Liu, M., Zhang, T.. A recognition method of soybean leaf diseases based on an improved deep learning model. Front. Plant Sci., 2023, 13, 878834.
14. Deb M., Garai A., Das A. et al.. LS-Net: a convolutional neural network for leaf segmentation of rosette plants. Neural Comput. & Applic., 2022, 34: 18511–18524.
15. Fang, J., Liu, H., Zhang, L., Liu, J., Liu, H.. Region-edge-based active contours driven by hybrid and local fuzzy region-based energy for image segmentation. Inf. Sci., 2021, 546 (6): 397-419.
16. Ngugi, L.C., Abdelwahab, M., Abo-Zahhad, M.. A new approach to learning and recognizing leaf diseases from individual lesions using convolutional neural networks. Information Processing in Agriculture, 2021, 10(1): 11-27.
17. Liu, B. L.. Research on the segmentation method of rice leaf disease image. Appl. Mech. Mater., 2012, 223: 1339-1344.
18. Kalaivani, S., Shantharajah, S., Padma, T.. Agricultural leaf blight disease seg mentation using indices based histogram intensity segmentation approach. Multimedia Tools Appl., 2020, 79 (13): 9145-9159.
19. Fang, J., Liu, H., Liu, J., Zhang, L., Liu, H.. Fuzzy region-based active contour driven by global and local fitting energy for image segmentation. Applied Soft Comp., 2021, 200, 106982:1-16.
20. Jothiaruna, N., Joseph Abraham Sundar, K., Ifjaz Ahmed, M.. A disease spot segmentation method using comprehensive color feature with multi-resolution channel and region growing. Multimedia Tools Appl., 2021, 80 (3): 3327–3335.
21. Xiong, L., Zhang, D., Li, K., Zhang, L.. The extraction algorithm of color disease spot image based on Otsu and watershed. Soft Comput., 2020, 24 (10): 7253-7263.
22. Tian, K., Li, J., Zeng, J., Evans, A., Zhang, L.. Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm. Comput. Electron. Agric., 2019, 165, 104962.

 doi:10.20944/preprints202405.1584.v1

23. Picon, A., San-Emeterio, M. G., Bereciartua-Perez, A., Klukas, C., Eggers, T., NavarraMestre, R.. Deep learning-based segmentation of multiple species of weeds and corn crop using synthetic and real image datasets. Comput. Electron. Agric., 2022, 194, 106719.
24. Long, J., Shelhamer, E., Darrell, T.. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp 3431-3440.
25. Ronneberger, O., Fischer, P., Brox, T.. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234-241.
26. Oktay, O., Schlemper, J., Folgoc, L.L., et al.. Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv: 1804.03999, 2018.
27. Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J.. UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging, 2019, 39 (6): 1856-1867.
28. Li, H.，Xiong, P., An, J., Wang L.. Pyramid attention network for semantic segmentation. arXiv preprint arXiv: 1805.10180, 2018.
29. Chen, J., Lu, Y., Yu, Q., et al.. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.
30. Fang, J., Jiang, H., Zhang, S., etal.. BAF-Net: Bidirectional attention fusion network via CNN and transformers for the pepper leaf segmentation. Front. Plant Sci., 2023, 14, 1123410.
31. Zhang, S., Zhao X., Tian Q.. Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM. IEEE T. Affect. Comput., 2022, 13(2): 680-688.
32. Zhang, S., Yang, Y., Chen, C., et al.. Multimodal emotion recognition based on audio and text by using hybrid attention networks. Biomed. Signal Process, 2023, 85, 105052.
33. Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., et al.. MLP-Mixer: An all-mlp architecture for vision. arXiv preprint arXiv: 2105.01601, 2021.
34. Ding, X. H., Chen, H. H., Zhang, X.Y., Han, J.G., Ding, G.G.. RepMLPNet: Hierarchical Vision MLP with Re-parameterized Locality. arXiv preprint arXiv: 2112.11081, 2022
35. Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., Asari V. K.. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation, arXiv preprint arXiv:1802.06955, 2018.
36. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.. Deformable convolutional networks. arXiv preprint arXiv: 1703.06211, 2017.
37. Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., Su, R.. DUNet: A deformable network for retinal vessel segmentation. Knowledge-Based Systems, 2019, 178: 149-162.
38. Xiang, T., Zhang, C., Liu, D., Song, Y., Huang, H., and Cai, W.. BiO-Net: Learning recurrent bi-directional connections for encoder-decoder architecture. arXiv preprint arXiv: 2007.00243, 2020.
39. Hu, J., Shen, L., Sun, G.. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.
40. Li, B., Kang, G., Cheng, K., Zhang N.. Attention-guided convolutional neural network for detecting pneumonia on chest x-rays. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 4851-4854.
41. Fan, D. P., Ji, G. P., Zhou, T., et al.. Pranet: Parallel reverse attention network for polyp segmentation. arXiv preprint arXiv: 1802.06955, 2020.
42. Devlin, J., Chang, M.-W., Lee, K., et al.. Bert: Pre-training of deep bidirectional Transformers for language understanding. ArXiv Preprint ArXiv: 1810.04805, 2018.
43. Melas-Kyriazi, L.. Do you even need attention a stack of feed-forward layers does surprisingly well on imagenet. arXiv preprint arXiv: 2105.02723, 2021.
44. Liu, H., Dai, Z., So, D., and Le, Q. V.. Pay attention to MLPs. arXiv preprint arXiv: 2105.08050, 2021.
45. Chen, S., Xie, E., Ge, C., Liang, D., Luo, P.. CycleMLP: A MLP-like architecture for dense prediction, arXiv preprint arXiv:2107.10224, 2021.
46. Tu, Z., Talebi, H., Zhang, H. F., Yang, Milanfar, P., Bovik, A., and Li, Y.. Maxim: Multi-axis mlp for image processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5769-5780.
47. Valanarasu, J. M. J., Patel, V. M.. UNeXt: MLP-based rapid medical image segmentation network. arXiv preprint arXiv: 2203.04967, 2022.

 doi:10.20944/preprints202405.1584.v1

48. Lv, J., Hu Y., Fu, Q., Zhang, Z. , Hu, Y., Lv, L., Yang, G., Li, J., Zhao, Y.. CM-MLP: Cascade multi-scale MLP with axial context relation encoder for edge segmentation of medical image. arXiv preprint arXiv: 2208.10701, 2022.
49. Tang, H., Liu, X., Sun, S., Yan, X., and Xie, X.. Recurrent mask refinement for few-shot medical image segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3918-3928.