CAPBNet: Channel-Attentive Pyramid and Bottleneck-Guided Context Network for Plant Disease Segmentation

Xingyu Ren[1], Rui Teng[1(✉)], Jinlai Zhang[2(✉)], and Yuanhao Yang[2]

[1] Elite Engineering School, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China

Abstract. Accurate plant disease segmentation remains challenging due to complex natural backgrounds, diverse lesion appearances, and ambiguous disease boundaries. In this paper, we propose a Channel-Attentive Pyramid and Bottleneck-Guided Context Network for plant disease segmentation. The proposed network improves DeepLabV3-ResNet50 by recalibrating multi-scale atrous branch features, refining lesion-related spatial context, and combining pixel-wise supervision with region-level overlap optimization. Experiments on the PlantSeg115 dataset show that the proposed method achieves 44.52% mIoU and 57.51% mAcc, outperforming the DeepLabV3 baseline by 1.43 and 2.11 percentage points, respectively. These results validate the effectiveness of channel-attentive multi-scale representation and bottleneck-guided context refinement for plant disease semantic segmentation.

Keywords: Plant disease segmentation, Semantic segmentation, Channel-attentive pyramid, Bottleneck-guided context refinement.

1 Introduction

Plant disease segmentation is an important task in agricultural computer vision, with practical value for disease monitoring, precision management, and intelligent crop protection. Deep learning has been widely investigated in agricultural vision and plant disease diagnosis, showing strong potential for automatic crop monitoring and disease assessment [15,7]. Compared with image-level disease recognition, semantic segmentation provides pixel-level localization of infected regions, enabling more detailed analysis of lesion distribution, disease severity, and spatial characteristics. This is particularly useful for plant disease diagnosis in natural scenes, where accurate lesion boundaries can provide direct visual evidence for subsequent disease assessment and decision-making.

Existing studies on plant disease analysis have progressed from image-level recognition to pixel-level segmentation. Early plant disease datasets and recognition methods mainly focused on classifying disease categories from whole images [14], [19], [24], [20]. With the development of dense prediction models, plant disease segmentation has received increasing attention because it can provide more precise lesion localization and support fine-grained disease assessment [21,29,6]. Meanwhile, general semantic segmentation models, such as fully convolutional networks, encoder-decoder architectures, and multi-scale context networks, have provided effective technical foundations

for dense visual prediction [17], [1], [22], [32]. Among them, DeepLabV3 is a representative framework that introduces Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale contextual features through parallel atrous convolution branches [4].

Despite these advances, accurate plant disease segmentation in natural scenes remains challenging. First, plant disease images often contain complex backgrounds, illumination variations, leaf occlusion, shadows, and irrelevant texture interference. These factors make it difficult for segmentation models to distinguish disease regions from healthy tissues and background areas. Second, disease lesions usually exhibit large variations in scale, shape, color, and texture. Some lesions occupy large continuous regions, while others appear as small and scattered spots. Third, the boundaries between diseased and healthy regions are often ambiguous, especially when symptoms are weak, low-contrast, or gradually distributed. These challenges may lead to incomplete lesion masks, rough boundaries, and background false positives.

Although DeepLabV3 can effectively enlarge the receptive field through ASPP, its original multi-branch design still has limitations for plant disease segmentation. Different atrous branches may contribute unequally to different lesion patterns, but the original ASPP structure treats branch features in a relatively uniform manner after multi-scale extraction. As a result, directly concatenating multi-scale features may introduce redundant or weakly relevant responses, especially in complex disease scenes. In addition, after multi-branch feature fusion, the segmentation head still needs to further refine lesion-related spatial context. Without effective spatial refinement, the predicted masks may suffer from fragmented regions, inaccurate boundaries, or false responses caused by leaf veins, shadows, and background textures.

To address these problems, we propose CAPBNet, a Channel-Attentive Pyramid and Bottleneck-Guided Context Network for plant disease segmentation. The proposed framework is built upon the DeepLabV3-ResNet50 baseline and enhances the segmentation head from both multi-scale representation and spatial context perspectives. Specifically, a Channel-Attentive Pyramid (CAP) Block is introduced to recalibrate multi-scale atrous branch features before feature fusion, allowing the model to adaptively emphasize informative lesion-related channels from different receptive fields. Furthermore, a Bottleneck-Guided Context Attention (BCA) Block is designed to refine the fused feature representation by generating a lesion-related spatial attention response and enhancing disease-relevant regions through residual fusion. In training, Pixel-Region Consistency Optimization (PRCO) is adopted to balance pixel-wise category discrimination and region-level overlap consistency, thereby improving the completeness of predicted lesion masks without introducing additional inference cost.

Experiments are conducted on the PlantSeg115 dataset, which contains 115 plant disease categories and pixel-level annotations. The proposed CAPBNet achieves 44.52% mIoU and 57.51% mAcc. Compared with the DeepLabV3 baseline, CAPBNet improves mIoU by 1.43 percentage points and mAcc by 2.11 percentage points. These results demonstrate that channel-attentive multi-scale representation and bottleneck-guided context refinement are effective for improving plant disease segmentation under complex natural conditions.

The main contributions of this paper are summarized as follows:

• We propose CAPBNet, an enhanced DeepLabV3-based framework for plant disease segmentation, which improves lesion representation through channel-attentive pyramid modeling and bottleneck-guided context refinement.

• We design a CAP Block to recalibrate multi-scale atrous branch features, enabling the model to adaptively emphasize informative channels from different receptive fields before feature fusion.

• We introduce a BCA Block to refine lesion-related spatial context after multi-scale feature fusion, improving the consistency of predicted disease regions and reducing background interference.

The remainder of this paper is organized as follows. Section 2 reviews related work on plant disease segmentation, semantic segmentation networks, and attention-based feature refinement. Section 3 introduces the proposed CAPBNet framework and describes the CAP Block, BCA Block, and PRCO strategy in detail. Section 4 presents the experimental setup, quantitative comparison, ablation study, and qualitative analysis. Section 5 concludes this paper.

2 Related Work

2.1 Plant Disease Segmentation

Plant disease analysis has become an important research topic in agricultural computer vision [14], [19], [24], [20]. Early studies mainly focused on image-level disease classification [33-37], where the goal is to assign a disease category to the whole image. Existing studies have also shown that dataset scale, image diversity, and deep model design strongly affect the robustness of plant disease recognition systems [2], [23]. Although these methods can provide useful diagnostic information, they usually cannot indicate the exact location and shape of diseased regions. For practical agricultural applications, pixel-level segmentation is more informative because it can localize lesion areas and support more detailed analysis of disease severity and spatial distribution [21], [29], [6].

With the development of deep learning, convolutional neural networks have been widely used for plant disease segmentation. Encoder-decoder networks and fully convolutional architectures can learn hierarchical visual representations and produce dense predictions. These methods improve segmentation performance compared with traditional hand-crafted feature methods. However, plant disease images collected in natural scenes often contain complex backgrounds, illumination variations, occlusion, and large intra-class appearance changes. Disease regions may also vary significantly in scale, color, texture, and shape. Therefore, segmentation models still face difficulties in distinguishing small lesions from background noise and in preserving complete lesion boundaries [21], [29].

Another challenge lies in the fine-grained nature of plant disease segmentation. Different diseases may share similar visual symptoms, while the same disease may appear differently across plant species or growth stages. This increases the demand for discriminative feature representation. A model needs to capture both local lesion details and broader semantic context. Methods that rely only on local convolutional [38-45] responses may fail to model sufficient contextual information, while methods that emphasize global semantics may lose small or weak disease regions. Therefore, effective plant disease segmentation requires a better balance between local detail representation and high-level semantic understanding [29], [6].

2.2 Semantic Segmentation with Multi-scale Context

Semantic segmentation aims to assign a semantic category to each pixel in an image. Since the introduction of fully convolutional networks, deep segmentation models have become the dominant paradigm for dense prediction by replacing image-level classification heads with pixel-level prediction structures [17]. Encoder-decoder architectures further improve segmentation by recovering spatial details from high-level semantic features. Representative models such as SegNet and U-Net exploit decoder structures or skip connections to recover fine-grained spatial information and produce dense segmentation maps [1], [22].

Multi-scale context modeling is another important direction in semantic segmentation. PSPNet introduces pyramid pooling to aggregate contextual information from different spatial regions, which improves the representation of objects and regions with diverse scales [32]. The DeepLab series models further explore atrous convolution for dense prediction. DeepLabV3 adopts Atrous Spatial Pyramid Pooling (ASPP), where parallel atrous convolution branches with different dilation rates are used to capture multi-scale contextual information while maintaining feature resolution [4]. DeepLabV3+ extends this framework with an encoder-decoder structure to refine object boundaries and recover low-level spatial details [5]. In addition, UPerNet integrates pyramid pooling and feature pyramid representations to exploit hierarchical multi-level features for scene understanding [31].

The multi-scale design of DeepLabV3 is suitable for plant disease segmentation because disease lesions often appear with diverse sizes and irregular shapes. Small scattered spots and large infected regions may require different receptive fields for effective representation. However, the original ASPP structure mainly extracts multi-scale features and fuses them in a relatively direct manner. It does not explicitly evaluate the importance of different feature channels from each atrous branch. In complex plant disease scenes, some branch features may contain useful lesion cues, while others may include redundant background responses. Direct fusion of all branch outputs may therefore weaken discriminative disease information and introduce irrelevant interference.

To address this limitation, recent segmentation studies have increasingly explored adaptive feature refinement and attention-based recalibration for improving multi-scale representations. For plant disease segmentation, such mechanisms are particularly useful because the contribution of different scales may vary according to lesion size, texture, and surrounding background. Therefore, a channel-attentive pyramid structure can provide a more flexible way to refine multi-scale atrous features before feature fusion.

2.3 Context Refinement and Loss Optimization

In addition to multi-scale feature extraction, context refinement is also important for semantic segmentation. Disease regions are often irregular, fragmented, and ambiguous near boundaries. After multi-scale feature fusion, the segmentation head still needs to refine spatial context to improve the consistency of predicted regions. Spatial context modeling can help the model focus on lesion-related areas and reduce false responses caused by background texture, leaf veins, shadows, or other visual noise.

Attention mechanisms have been widely studied to improve feature representation in deep neural networks. Squeeze-and-Excitation networks introduce channel recalibration to adaptively emphasize informative feature channels [12]. CBAM further com-

bines channel attention and spatial attention to refine feature responses from complementary perspectives [30]. ECA-Net improves channel attention with efficient local cross-channel interaction, avoiding dimensionality reduction and introducing only limited additional complexity [27]. These channel-oriented attention mechanisms provide useful inspiration for recalibrating multi-scale branch features in segmentation networks.

Spatial and contextual attention mechanisms are also important for dense prediction. Non-local networks capture long-range dependencies by modeling relationships between distant positions in feature maps [28]. CCNet improves semantic segmentation by using criss-cross attention to efficiently aggregate contextual information along horizontal and vertical directions [13]. DANet introduces position attention and channel attention to enhance scene segmentation by jointly modeling spatial dependencies and channel relationships [8]. For plant disease segmentation, these context refinement strategies are useful because lesions are not only defined by color or texture at a single pixel, but also by their surrounding structure and spatial distribution.

Loss function design also plays an important role in segmentation performance. Cross-entropy loss is commonly used because it provides stable pixel-wise category supervision. It is effective for learning discriminative class boundaries and is especially important in multi-class segmentation tasks. However, cross-entropy mainly optimizes pixel-wise classification and may not directly align with region-level metrics such as intersection-over-union. Lovasz-Softmax provides a tractable surrogate for directly optimizing the intersection-over-union measure in neural networks, making it useful for improving region-level mask consistency [3].

Therefore, combining pixel-wise classification supervision with region-level overlap optimization is a reasonable strategy for plant disease segmentation. Pixel-wise supervision can maintain stable category discrimination, while region-oriented optimization can encourage better mask-level consistency. Based on these observations, the proposed framework integrates channel-attentive pyramid modeling, bottleneck-guided context refinement, and Pixel-Region Consistency Optimization to improve plant disease segmentation under complex natural conditions.

3 Methodology

3.1 Overview of CAPBNet

In this section, we introduce the proposed CAPBNet, a Channel-Attentive Pyramid and Bottleneck-Guided Context Network for plant disease segmentation. The framework is built upon the DeepLabV3-ResNet50 baseline and aims to improve lesion representation from three aspects: multi-scale atrous feature recalibration, spatial context refinement, and segmentation-oriented loss optimization.

Given an input plant disease image, the ResNet-50-V1c backbone first extracts hierarchical feature representations. ResNet provides a residual learning framework that facilitates the optimization of deep convolutional networks and has been widely used as a strong backbone for dense prediction tasks [10]. The shallow layers contain low-level visual cues such as edges and textures, while the deeper layer provides high-level semantic information. Following the DeepLabV3-style design, the main segmentation head uses the high-level feature from the last backbone stage as the input for dense prediction. This feature is then passed into the Channel-Attentive Pyramid (CAP)

Block, which enhances multi-scale atrous features through branch-wise channel recalibration. After multi-scale fusion, the Bottleneck-Guided Context Attention (BCA) Block further refines lesion-related spatial context and produces a more discriminative representation for final classification.

The overall prediction process can be summarized as follows. The backbone extracts the high-level feature F_4 from the input image. The CAP Block transforms F_4 into a multi-scale recalibrated feature F_{cap}. Then, the BCA Block refines F_{cap} and generates the enhanced feature F_{bca}. Finally, a 1×1 classification layer maps the refined feature into 116-class segmentation logits, and the final mask is obtained by pixel-wise category prediction.

During training, the main segmentation head is supervised by the PRCO objective, while an auxiliary head is used to provide additional supervision to intermediate backbone features. The auxiliary head is only used during training and is removed during inference. In this way, CAPBNet keeps the original efficiency of the DeepLabV3-style inference pipeline while improving multi-scale representation and spatial context modeling. The overall architecture of CAPBNet is shown in Figure 1.

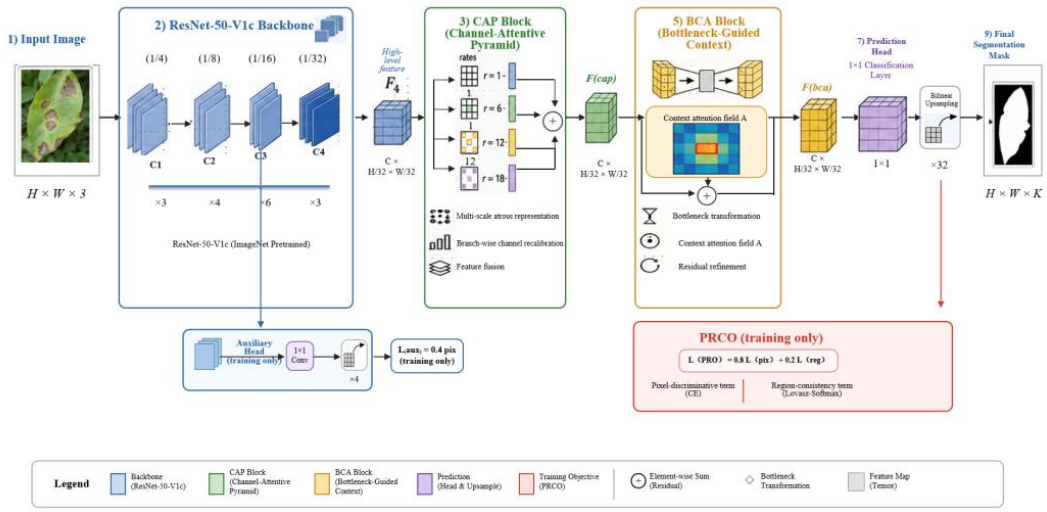

Fig. 1. Overall architecture of the proposed CAPBNet. The framework is built upon DeepLabV3-ResNet50 and enhances the segmentation head with the Channel-Attentive Pyramid (CAP) Block and the Bottleneck-Guided Context Attention (BCA) Block.

3.2 Channel-Attentive Pyramid Block

Plant disease lesions usually show large variations in scale. Some disease regions appear as small and scattered spots, while others occupy large continuous areas. Therefore, multi-scale contextual modeling is important for accurate lesion segmentation. Multi-branch convolutional designs have been widely used to capture complementary receptive fields, and feature pyramid representations further demonstrate the importance of multi-scale feature learning in visual recognition tasks [26], [16]. The orig-

inal DeepLabV3 head uses atrous spatial pyramid pooling to extract features from different receptive fields. However, different atrous branches may not contribute equally to different lesion patterns. Directly concatenating all branch outputs may introduce redundant background responses and weaken useful lesion-related cues.

To address this limitation, we introduce the Channel-Attentive Pyramid (CAP) Block. The CAP Block follows the multi-branch atrous pyramid structure, but inserts a channel recalibration operation after each atrous branch. In this way, each branch can adaptively emphasize informative channels before feature fusion.

Let F_4 denote the high-level feature extracted by the backbone. The multi-scale atrous branches are defined in Eq. (1):

$$B_m = \phi_m(F_4), \quad m \in \{1,2,3,4\}, \tag{1}$$

where $\phi_m(\cdot)$ denotes the convolutional transformation of the m-th atrous branch. These branches correspond to different receptive fields, including local, small-scale, middle-scale, and large-scale contextual patterns. In addition, an image-level pooling branch is used to provide global contextual information.

For each branch feature B_m, global average pooling is first used to obtain a compact channel descriptor in Eq. (2):

$$z_m^c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} B_m^c(i,j), \tag{2}$$

where $B_m^c(i,j)$ represents the response of the c-th channel at spatial position (i,j), and H and W denote the spatial size of the feature map. Then, a lightweight one-dimensional convolution is applied to model local cross-channel interaction as shown in Eq. (3):

$$s_m = \sigma\big(\mathrm{Conv1D}_k(z_m)\big), \tag{3}$$

where s_m is the channel attention weight of the m-th branch, $\sigma(\cdot)$ denotes the sigmoid activation function, $\mathrm{Conv1D}_k(\cdot)$ denotes one-dimensional convolution with kernel size k, and k is the adaptive kernel size. The recalibrated branch feature is obtained by Eq. (4):

$$\tilde{B}_m = B_m \odot s_m, \tag{4}$$

where $\odot$ denotes channel-wise multiplication.

After branch-wise channel recalibration, all enhanced branch features are concatenated and fused by a 1×1 convolution as shown in Eq. (5):

$$F_{cap} = \mathrm{Conv}_{1\times1}\big([\tilde{B}_1, \tilde{B}_2, \tilde{B}_3, \tilde{B}_4, B_{pool}]\big). \tag{5}$$

Here, B_{pool} denotes the image-level pooling branch. Through this design, the CAP Block can preserve useful multi-scale lesion cues and suppress redundant channel responses before feature fusion. This is beneficial for plant disease segmentation because lesion regions may appear at different scales and may require different channel responses from different atrous branches. The detailed structure of the CAP Block is illustrated in Figure 2.

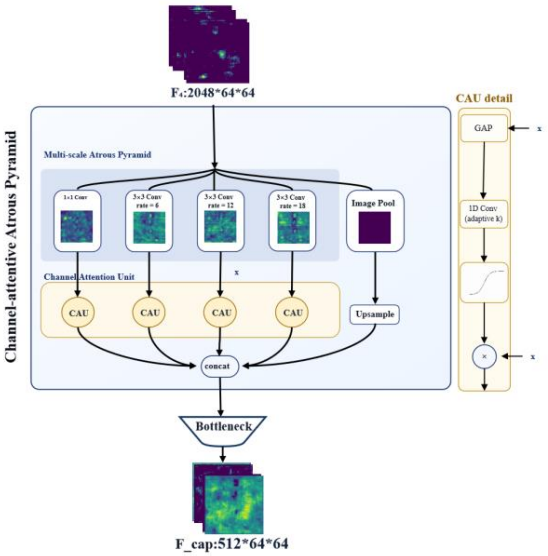

Fig. 2. Structure of the Channel-Attentive Pyramid (CAP) Block. Each atrous branch is followed by channel recalibration, enabling the model to adaptively emphasize informative multi-scale lesion features before feature fusion.

3.3 Bottleneck-Guided Context Attention Block

Although the CAP Block enhances multi-scale channel representation, the fused feature may still contain spatially redundant responses caused by complex backgrounds, leaf veins, shadows, and texture interference. In plant disease segmentation, lesion regions usually exhibit irregular shapes and ambiguous boundaries. Therefore, after multi-scale feature fusion, it is necessary to further refine lesion-related spatial context before the final pixel-wise classification.

To address this problem, we introduce the Bottleneck-Guided Context Attention (BCA) Block. The BCA Block takes the fused feature F_{cap} as input and refines it through bottleneck transformation, local context aggregation, directional depth-wise convolution, and residual gating. Direction-aware spatial modeling has been shown to be effective for pixel-wise prediction, where long and narrow kernels such as $1 \times N$ and $N \times 1$ can capture contextual dependencies along different spatial directions [11]. In addition, recent convolutional attention designs indicate that efficient convolutional operations can provide effective contextual encoding for semantic segmentation [9]. Different from directly replacing the fused feature, BCA enhances lesion-related responses in a residual manner, which helps preserve the original multi-scale representation while suppressing irrelevant background responses.

Given the fused feature F_{cap}, a bottleneck transformation is first applied in Eq. (6):

$$F_b = \psi(F_{cap}), \tag{6}$$

where $\psi(\cdot)$ denotes the bottleneck convolutional transformation. This step provides a compact feature representation for subsequent context refinement. Then, local spatial context is aggregated by a 7×7 average pooling operation in Eq. (7):

$$F_p = \mathrm{AvgPool}_{7\times7}(F_b). \tag{7}$$

To capture directional spatial dependencies, horizontal and vertical depth-wise convolutions are further applied to the pooled feature as shown in Eq. (8):

$$F_d = \mathrm{Conv}_{1\times1}\left(\mathrm{DWConv}_{11\times1}\left(\mathrm{DWConv}_{1\times11}\left(\mathrm{Conv}_{1\times1}(F_p)\right)\right)\right), \tag{8}$$

where $\mathrm{DWConv}_{1\times11}(\cdot)$ and $\mathrm{DWConv}_{11\times1}(\cdot)$ model horizontal and vertical contextual dependencies, respectively. The two 1×1 convolutions are used for feature transformation before and after directional context modeling. The obtained F_d represents the directional context response generated from the bottleneck feature.

Finally, sigmoid-gated residual fusion is used to obtain the refined feature in Eq. (9):

$$F_{bca} = F_{cap} + F_{cap} \odot \sigma(F_d), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $\odot$ denotes element-wise multiplication. In this formulation, $\sigma(F_d)$ serves as the spatial gating response for enhancing lesion-related regions. The residual connection preserves the original fused representation, while the gated branch strengthens disease-relevant spatial responses.

This design is beneficial for plant disease segmentation because disease regions often contain weak boundaries, irregular shapes, and background interference. By combining bottleneck transformation, local context aggregation, directional depth-wise convolution, and residual gating, the BCA Block can refine lesion-related spatial context without discarding the original multi-scale feature. As a result, the segmentation head can generate more spatially consistent and robust lesion representations before final classification. The detailed structure of the BCA Block is shown in Figure 3.

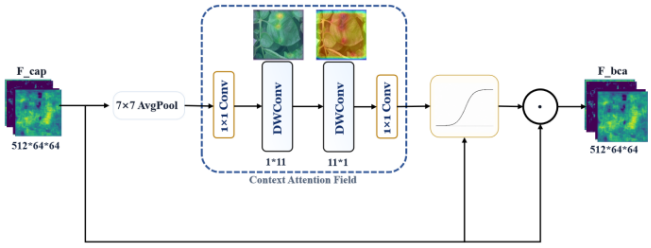

Fig. 3. Structure of the Bottleneck-Guided Context Attention (BCA) Block. The fused feature is refined through bottleneck transformation, local context aggregation, directional depth-wise convolution, and sigmoid-gated residual fusion.

3.4 Pixel-Region Consistency Optimization

In semantic segmentation, the training objective directly affects pixel-wise category discrimination and region-level mask consistency. Region-level overlap objectives have been widely studied for segmentation tasks with foreground-background imbalance, since they can directly encourage better agreement between predicted masks and ground-truth regions [25]. For plant disease segmentation, this issue is particularly important because disease lesions often have irregular shapes, ambiguous boundaries, and large intra-class appearance variations. A conventional pixel-wise objective can provide stable category supervision, but it may not sufficiently encourage the structural completeness and overlap consistency of predicted lesion masks.

To address this problem, we introduce a Pixel-Region Consistency Optimization (PRCO) strategy for CAPBNet. Instead of relying on a single optimization criterion,

PRCO formulates a unified Pixel-Region Objective, denoted as $\mathcal{L}_{\mathrm{PRO}}$, to jointly consider pixel-level discriminative learning and region-level mask consistency. The objective consists of a pixel-discriminative term and a region-consistency term, as defined in Eq. (10):

$$L_{PRO} = \omega_{pix} L_{pix} + \omega_{reg} L_{reg}, \quad \omega_{pix} + \omega_{reg} = 1, \tag{10}$$

where $\mathcal{L}_{\mathrm{pix}}$ focuses on pixel-level category discrimination, $\mathcal{L}_{\mathrm{reg}}$ encourages region-level overlap consistency, and ω_{pix} and ω_{reg} are the corresponding balancing coefficients.

Let Ω denote the set of valid pixels, C denote the number of semantic classes, z_i^c denote the logit of pixel i for class c, and y_i denote the ground-truth label of pixel i. The predicted probability of pixel i belonging to class c is computed by the softmax function in Eq. (11):

$$p_i^c = \frac{\exp(z_i^c)}{\sum_{k=1}^{C} \exp(z_i^k)}. \tag{11}$$

The pixel-discriminative term is defined in Eq. (12):

$$L_{pix} = -\frac{1}{|\Omega|} \sum_{i \in \Omega} \log p_i^{y_i}. \tag{12}$$

This term provides stable pixel-wise supervision and helps the model distinguish fine-grained disease categories.

To further improve mask-level consistency, the region-consistency term is introduced from an IoU-oriented perspective. For class c, the binary ground-truth indicator is defined in Eq. (13):

$$y_i^c = \mathbb{I}(y_i = c), \tag{13}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The per-pixel margin error for class c is formulated in Eq. (14):

$$m_i^c = \begin{cases} 1 - p_i^c, & \text{if } y_i^c = 1, \\ p_i^c, & \text{if } y_i^c = 0, \end{cases} \tag{14}$$

Let $m_{(1)}^c \geq m_{(2)}^c \geq \cdots \geq m_{(|\Omega|)}^c$ be the sorted errors in descending order, and let $y_{(i)}^c$ be the corresponding sorted binary labels. The region-consistency term is computed in Eq. (15):

$$L_{reg} = \frac{1}{|C_v|} \sum_{c \in C_v} \sum_{i=1}^{|\Omega|} m_{(i)}^c \, g_i^c, \tag{15}$$

where $\mathcal{C}_v$ denotes the set of valid classes and g_i^c is the discrete gradient of the Jaccard error after sorting, as shown in Eq. (16):

$$g_i^c = J_i^c - J_{i-1}^c, \quad J_0^c = 0. \tag{16}$$

Here, J_i^c is defined in Eq. (17):

$$J_i^c = 1 - \frac{G_c - \sum_{j=1}^{i} y_{(j)}^c}{G_c + \sum_{j=1}^{i} \left(1 - y_{(j)}^c\right)}, \quad G_c = \sum_{j=1}^{|\Omega|} y_j^c. \tag{17}$$

In this way, $\mathcal{L}_{\mathrm{reg}}$ provides region-level supervision by approximating the optimization of the Jaccard error, which is closely related to the mIoU evaluation metric.

In our implementation, ω_{pix} and ω_{reg} are set to 0.8 and 0.2, respectively. This setting maintains stable pixel-wise category learning while introducing region-level overlap guidance for more complete lesion masks. In addition, an auxiliary segmentation head is used during training to provide intermediate supervision for the backbone feature representation as shown in Eq. (18):

$$L_{aux} = \lambda L_{pix}^{aux}, \tag{18}$$

where $\mathcal{L}_{\text{pix}}^{\text{aux}}$ denotes the auxiliary pixel-discriminative loss and λ is set to 0.4. Therefore, the final training objective is formulated in Eq. (19):

$$L_{total} = L_{PRO} + L_{aux}. \tag{19}$$

Through this optimization design, PRCO provides complementary supervision for plant disease segmentation. The pixel-discriminative term improves category-level discrimination among multiple disease classes, while the region-consistency term encourages the predicted masks to better match ground-truth lesion regions. The auxiliary supervision further improves feature learning during training and is removed during inference. Therefore, PRCO improves segmentation-oriented optimization without introducing additional inference cost.

4 Experiments

4.1 Experiment Setup

All experiments were implemented on a server equipped with 8 NVIDIA Tesla P40 GPUs, each with 24 GB memory. Four GPUs were used for model training. The batch size was set to 4 on each GPU, resulting in a total batch size of 16. The models trained under our implementation setting were optimized for 40k iterations. The stochastic gradient descent (SGD) optimizer was adopted with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate was set to 0.01 and decayed according to the polynomial learning rate schedule. For the DeepLabV3 baseline, cross-entropy loss was used as the basic optimization objective. For CAPBNet, the proposed PRCO strategy was adopted as the main training objective. The software environment was based on Python 3.8, PyTorch 1.12.0 with CUDA 11.3, and the MMSegmentation toolbox [18].

The experiments were conducted on the PlantSeg115 dataset [29], which is a large-scale in-the-wild plant disease segmentation dataset with pixel-level annotations. The dataset used in this work contains 11,458 images covering 115 plant disease categories. Following the provided data split, the dataset was divided into 7,916 training images, 1,247 validation images, and 2,295 test images. The segmentation task contains 116 semantic classes, including 115 disease classes and the background class. This setting is adopted consistently for all compared experiments to ensure that the reported results are evaluated under the same data protocol.

Mean Intersection over Union (mIoU) and Mean Accuracy (mAcc) were adopted as the evaluation metrics. The mIoU measures the average overlap between the predicted segmentation results and the ground-truth masks across all valid classes. It is defined in Eq. (20):

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

where N denotes the number of semantic classes, TP_i denotes the number of true positive pixels for the i-th class, FP_i denotes the number of false positive pixels for the i-th class, and FN_i denotes the number of false negative pixels for the i-th class.

The mAcc measures the mean of per-class pixel accuracy and is defined in Eq. (21):

$$\text{mAcc} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i}. \tag{21}$$

where N, TP_i, and FN_i have the same meanings as those in Eq. (20). A higher mIoU or mAcc indicates better segmentation performance. In this work, mIoU is used as the

primary evaluation metric, while mAcc is reported as a complementary indicator of per-class prediction accuracy.

4.2 Main Results

We compare CAPBNet with representative semantic segmentation methods on the PlantSeg115 dataset, including DeepLabV3, SegFormer, and STAR-Net. The reported quantitative results of SegFormer (MiT-B2), STAR-Net (Swin-T), and STAR-Net with Combo Loss are taken from the PlantSeg evaluation in the STAR-Net study [6], while DeepLabV3 and CAPBNet are evaluated under our implementation setting. The qualitative SegFormer (MiT-B2) results shown in Figures 4 and 8 are generated from our reproduced SegFormer model under the same visualization protocol. As shown in Table 1, CAPBNet achieves the best performance, with 44.52% mIoU and 57.51% mAcc. Compared with the DeepLabV3 baseline, CAPBNet improves mIoU by 1.43 percentage points and mAcc by 2.11 percentage points.

Compared with the reported PlantSeg results of SegFormer and STAR-Net, CAPBNet also achieves higher performance in both metrics. Specifically, CAPBNet outperforms SegFormer (MiT-B2) by 3.86 percentage points in mIoU and 2.27 percentage points in mAcc. It also outperforms STAR-Net (Swin-T) by 3.39 percentage points in mIoU and 4.84 percentage points in mAcc. These results indicate that CAPBNet provides competitive segmentation performance on PlantSeg115, especially for disease regions with diverse scales and complex visual appearances.

Table 1. Comparison with representative methods on the PlantSeg115 dataset.

Method	mIoU (%) ↑	mAcc (%) ↑
DeepLabV3	43.09	55.40
SegFormer (MiT-B2) [6]	40.66	55.24
STAR-Net (Swin-T, Combo Loss) [6]	39.34	51.95
STAR-Net (Swin-T) [6]	41.13	52.67
CAPBNet (Ours)	44.52	57.51

Figure 4 shows qualitative comparisons among DeepLabV3, SegFormer (MiT-B2), and CAPBNet on representative samples. CAPBNet produces more complete lesion masks and reduces missing responses in challenging disease regions, which is consistent with the design of the CAP Block and BCA Block. Figure 5 further reports category-wise foreground IoU on representative categories, showing that CAPBNet improves lesion localization across different disease appearances.

Figure 6 presents the training behavior of DeepLabV3 and CAPBNet. The losses decrease stably, while the decode accuracy curves show consistent optimization. Together with the validation comparison in Figure 7, these results indicate better validation performance of CAPBNet. Figure 8 provides further qualitative comparisons, where CAPBNet better preserves lesion completeness and reduces boundary errors.

Fig. 4. Qualitative comparison among DeepLabV3, SegFormer (MiT-B2), and CAPBNet on representative PlantSeg115 samples. From left to right, each row shows the ground-truth mask, the predictions of DeepLabV3, SegFormer (MiT-B2), and CAPBNet (Ours), respectively.

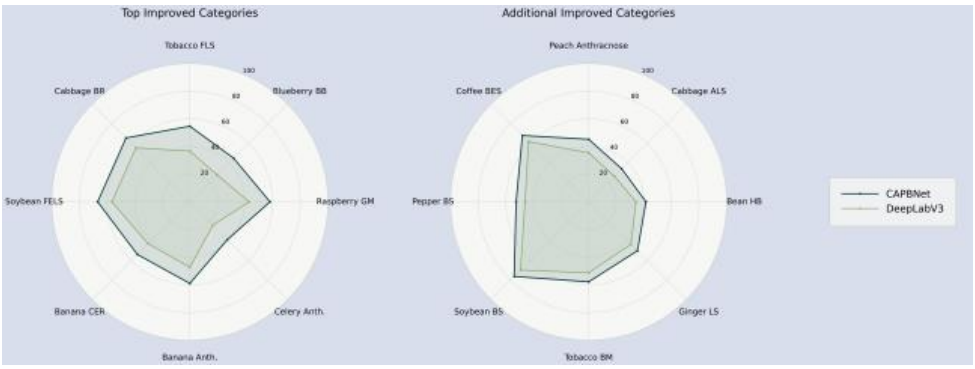

Fig. 5. Category-wise foreground IoU comparison between DeepLabV3 and CAPBNet on representative PlantSeg115 categories. The radial axis denotes foreground IoU in percentage, and the values are estimated from the saved qualitative comparison maps.

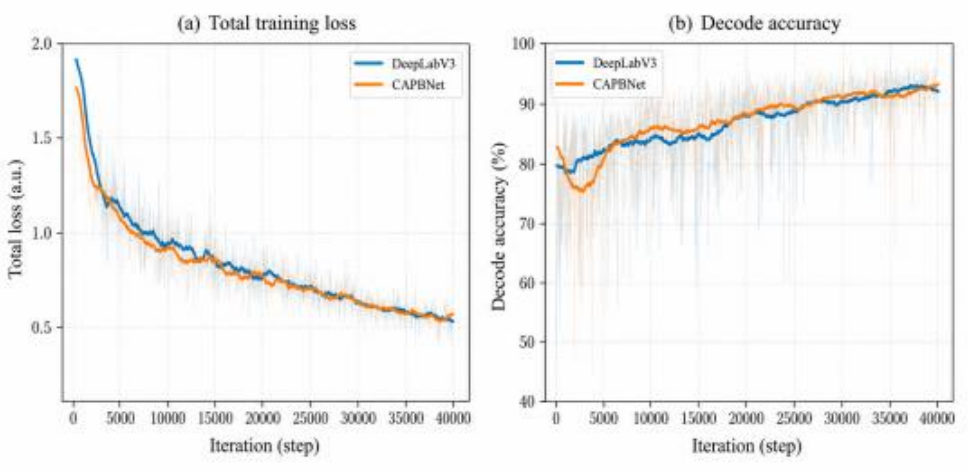

Fig. 6. Training curve comparison between DeepLabV3 and CAPBNet on the PlantSeg115 dataset. Subfigure (a) reports total training loss, and subfigure (b) reports decode accuracy.

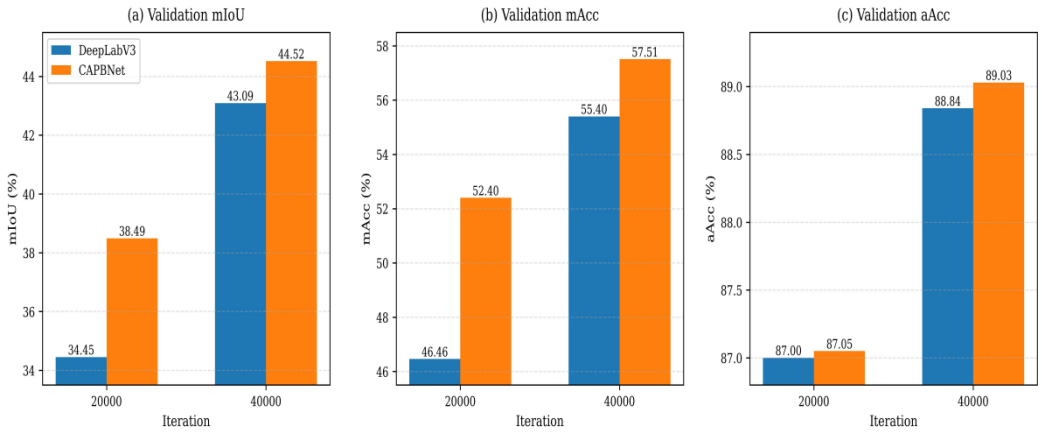

Fig. 7. Validation metric comparison between DeepLabV3 and CAPBNet at 20000 and 40000 iterations. From left to right, the subfigures report mIoU, mAcc, and aAcc, respectively.

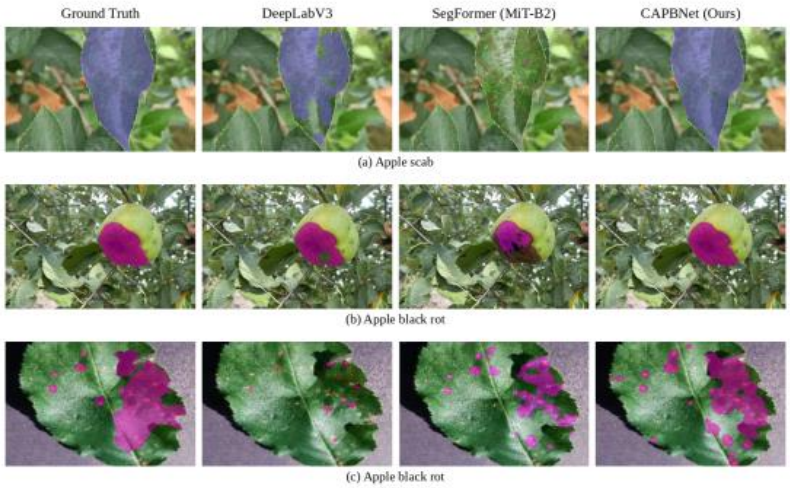

Fig. 8. Additional qualitative comparison among DeepLabV3, SegFormer (MiT-B2), and CAPBNet on representative PlantSeg115 samples.

4.3 Ablation Experiments

Table 2 studies the influence of different PRCO weight settings. The setting $\omega_{\text{pix}} = 0.80$ and $\omega_{\text{reg}} = 0.20$ achieves the best result, with 44.52% mIoU and 57.51% mAcc. When the region-consistency weight is increased to 0.25 or reduced to 0.15, the performance slightly decreases. This indicates that a proper balance between pixel-wise category discrimination and region-level overlap consistency is important for PlantSeg115 segmentation.

Table 3 reports the module-wise ablation results. Using the CAP Block alone obtains 43.15% mIoU and 55.79% mAcc. After adding the BCA Block, the performance increases to 44.04% mIoU and 56.34% mAcc, showing the benefit of bottleneck-guided context refinement. With the full design, CAPBNet achieves 44.52% mIoU and 57.51% mAcc, which confirms that CAP Block, BCA Block, and PRCO provide complementary improvements.

Table 2. Ablation study of PRCO weight settings on the PlantSeg115 dataset.

PRCO Weight Setting	mIoU (%) ↑	mAcc (%) ↑
ωpix = 0.75, ωreg = 0.25	43.85	56.88
ωpix = 0.80, ωreg = 0.20	44.52	57.51
ωpix = 0.85, ωreg = 0.15	43.84	56.97

Table 3. Module-wise ablation study on the proposed modules.

Method	mIoU (%) ↑	mAcc (%) ↑
DeepLabV3	43.09	55.40
CAP Block	43.15	55.79
CAP Block + BCA Block	44.04	56.34
CAP Block + BCA Block + PRCO	**44.52**	**57.51**

5 Conclusion

In this work, we present CAPBNet, an enhanced DeepLabV3-based framework for plant disease segmentation. The proposed method introduces channel-attentive pyramid modeling and bottleneck-guided context refinement to improve the representation of disease lesions under complex natural conditions. Specifically, the CAP Block strengthens multi-scale atrous features through branch-wise channel recalibration, enabling the segmentation head to better capture lesion patterns at different scales. The BCA Block further refines the fused feature representation by enhancing lesion-related spatial context and suppressing redundant background responses. In addition, PRCO balances pixel-wise category supervision and region-level overlap consistency, which helps the model generate more complete and stable segmentation masks.

Experiments on the PlantSeg115 dataset demonstrate the effectiveness of the proposed framework. Compared with the single CAP Block setting, introducing the BCA Block improves the segmentation performance from 43.15% to 44.04% in mIoU and from 55.79% to 56.34% in mAcc. Furthermore, combining CAP Block, BCA Block, and PRCO achieves 44.52% mIoU and 57.51% mAcc, showing that the proposed components provide complementary benefits for plant disease segmentation. These results indicate that CAPBNet can improve both lesion localization and region-level segmentation consistency.

Although CAPBNet improves the segmentation performance, there is still room for further research. Future work will focus on developing a more lightweight version of the proposed framework for practical agricultural deployment. In addition, we plan to evaluate the model on more plant species and real field scenarios to further verify its robustness and generalization ability. More effective strategies for small lesion segmentation and weak-boundary refinement will also be explored.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), 2481-2495 (2017).

2. Barbedo, J.G.A.: Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture 153, 46-53 (2018).

3. Berman, M., Rannen Triki, A., Blaschko, M.B.: The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4413-4421 (2018)

4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)

5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 801-818 (2018)

6. Fan, Y., Yu, M., Shen, L., Ma, J., Zeng, Z., Wang, H.: Robust plant disease segmentation in complex field environments: an in-depth analysis and validation with STAR-Net. Frontiers in Plant Science 16, 1706072 (2026).

7. Ferentinos, K.P.: Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture 145, 311-318 (2018).

8. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3146-3154 (2019)

9. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: SegNeXt: Rethinking convolutional attention design for semantic segmentation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 1140-1156 (2022)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770-778 (2016)

11. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4003-4012 (2020)

12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132-7141 (2018)

13. Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T.S.: CCNet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603-612 (2019)

14. Hughes, D.P., Salathé, M.: An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060 (2015)

15. Kamilaris, A., Prenafeta-Boldu, F.X.: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture 147, 70-90 (2018).

16. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 936-944 (2017).

17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431-3440 (2015)

18. MMSegmentation Contributors: MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. (2020)

19. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in Plant Science 7, 1419 (2016).

20. Moupojou, E., Tagne, A., Retraint, F., Tadonkengwa, A., Wilfried, D., Tapamo, H., Nkenlifack, M.: FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning. IEEE Access 11, 35398-35410 (2023).

21. Prashanth, K., Harsha, J.S., Kumar, S.A., Srilekha, J.: Towards accurate disease segmentation in plant images: A comprehensive dataset creation and network evaluation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7086-7094 (2024)

22. 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. pp. 234-241 (2015)

23. Saleem, M.H., Potgieter, J., Arif, K.M.: Plant disease detection and classification by deep learning. Plants 8(11), 468 (2019).

24. Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: PlantDoc: A dataset for visual plant disease detection. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. pp. 249-253 (2020).

25. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 240-248. Springer (2017).

26. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1-9 (2015).

27. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11534-11542 (2020)

28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794-7803 (2018)

29. Wei, T., Chen, Z., Yu, X., Chapman, S., Melloy, P., Huang, Z.: A large-scale in-the-wild dataset for plant disease segmentation. Scientific Data 13(205) (2026).

30. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3-19 (2018)

31. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision. pp. 418-434 (2018)

32. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881-2890 (2017)

33. Zhou S, Yuan Z, Yang D, et al. Pillarhist: A quantization-aware pillar feature encoder based on height-aware histogram, Proceedings of the computer vision and pattern recognition conference. 2025: 27336-27345.

34. Zhou S, Li L, Zhang X, et al. LiDAR-PTQ: Post-Training Quantization for Point Cloud 3D Object Detection, The Twelfth International Conference on Learning Representations. 2026.

35. Zhou, S., Cao, Y., Nie, J., Fu, Y., Zhao, Z., Lu, X., & Wang, S. Comptrack: Information bottleneck-guided low-rank dynamic token compression for point cloud tracking. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 40, No. 16, pp. 13773-13781. 2026

36. Zhou, H., Tang, J., Li, Y., Xiao, C., Hou, L., Ke, Z., & Yao, J. Comem: Compositional concept-graph memory for vision‐language adaptation. In The Fourteenth International Conference on Learning Representations. 2026.

37. Liu, H., Peng, N., Chen, Q., & Xiao, C. Affordance-first decomposition for continual learning in video-language understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3908-3919). 2026.

38. Xiao, C., Zhao, C., Ke, Z., & Shen, F. Curiosity-Driven Cooperation for Long-Tailed Multi-Label Learning. Neural Networks. 2026

39. Li, Z., Hu, Y., Chen, Z., Huang, Q., Qiu, G., Fu, Z., & Liu, M. (2026, March). Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 28, pp. 23373-23381).

40. Li, Z., Hu, Y., Chen, Z., Zhang, S., Huang, Q., Fu, Z., & Wei, Y. (2026, March). Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 8, pp. 6762-6770).

41. Li, Z., Chen, Z., Wen, H., Fu, Z., Hu, Y., & Guan, W. (2025, April). Encoder: Entity mining and modification relation binding for composed image retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 5, pp. 5101-5109).

42. Chen, Z., Hu, Y., Li, Z., Fu, Z., Wen, H., & Guan, W. (2025, October). Hud: Hierarchical uncertainty-aware disambiguation network for composed video retrieval. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 6143-6152).

43. He, Y., Zhang, C., Chen, F., & Cao, J. (2026). CineMatte: Background Matting for Virtual Production and Beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8725-8735).

44. Liu, T., Luo, Y. T., Pang, P. C. I., Zhang, H., Xiang, A., & Yang, Q. (2026). The role of multimodal generative AI in older adults' health management: Systematic scoping review. JMIR AI, 5(1), e84695.

45. Yang, J., Liu, T., Luo, Y. T., Niu, T., Pang, P., Xiang, A., & Yang, Q. (2025). Exploring the application boundaries of llms in mental health: A systematic scoping review. Frontiers in Psychology, 16, 1715306.