IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 March 2024, accepted 18 April 2024, date of publication 24 April 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393301

RESEARCH ARTICLE

RoI-Attention Network for Small Disease Segmentation in Crop Images

GOO-YOUNG MOON AND JONG-OK KIM, (Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Jong-Ok Kim (jokim@korea.ac.kr)

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food Agriculture and Forestry (IPET) through Open Field Smart Agriculture Technology Short-Term Advancement Program, funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA), under Grant 322033-3.

ABSTRACT We aim to contribute to deep learning based smart agriculture through semantic segmentation on crop images from real field environment. The key objective is the precise detection of diseases to facilitate the automation of agricultural management. The most significant issue is that the disease regions, serving as Regions of Interest (RoI), are small, making accurate prediction challenging. To address this issue, we propose a new framework of RoI-Attention Network (RA-Net) which additionally utilizes an RoI-attentive image that includes only regions predicted as disease and their surroundings from the input image. Using the RoI-attentive image, RA-Net enhances the representation power for disease regions by guiding the network to re-focus on RoI-associated context based on the initial prediction from the input. Using the proposed RoI-Attention stage, the coarse predictions of disease regions in crop images can be enhanced by incorporating additional sequential RoI-Attention and fusion stages. We have experimentally demonstrated the effectiveness of the proposed RA-Net in predicting small disease regions.

INDEX TERMS RoI-attention, semantic segmentation, small object detection, crop images, disease.

I. INTRODUCTION

Recently, there has been a growing interest in exploring smart agricultural approaches that employ advanced technologies like deep learning, sensors, and data analysis for the efficient cultivation and management of crops. Various studies have been conducted to analyze crop images through computer vision algorithms to measure or predict the degree of crop growth, or to detect the regions infected with diseases [1], [2], [3], [4], [5]. In this paper, we aim to contribute to automatic disease management by accurately segmenting the regions affected by diseases in crop images.

Segmentation, one of the key research topics in computer vision, enables the precise analysis of digital images by performing pixel-level classification. Recently, it has been widely used in various fields such as healthcare [6], autonomous driving [7], and smart agriculture. Segmentation for agriculture aims to accurately classify every pixel in a given crop image into three classes: plant, disease, and BG (background) [8], [9]. It is crucial to precisely detect disease regions for smart agricultural management, and thus, the regions of the disease class are designated as Regions of Interest (RoI) in crop images.

However, disease segmentation faces a critical challenge in accurately recognizing disease regions which are typically much smaller in size compared to the other plant and BG classes. Figure 1 shows the severe class imbalance between the disease and the other classes due to small disease regions. In Figure 1, the maize images on the left are from the public dataset [10], and the cabbage ones on the right were captured by our drone vehicles in the field. Especially, because crop images are mostly captured in a top-view perspective, as shown in Figure 1, the challenge of segmenting small disease regions can be even more prominent. This challenge can not only compromise the learning efficiency of a segmentation network but also pose significant obstacles in detecting disease regions, which are our main concern in this work. Such an issue is not confined to disease segmentation alone, as most of RoI classes are generally small. For instance, for the case of autonomous driving, it is crucial to accurately

The associate editor coordinating the review of this manuscript and approving it for publication was Tai Fei.

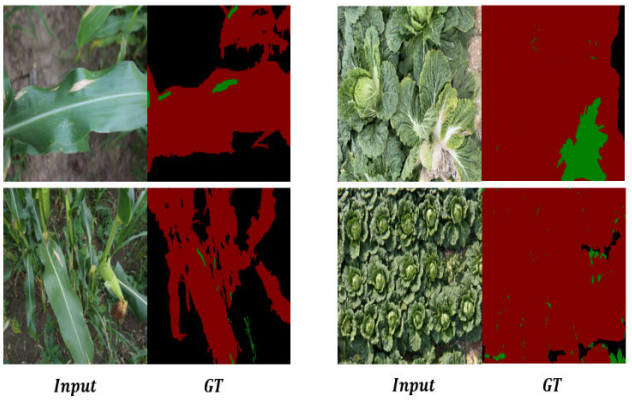

FIGURE 1. The examples of top-view crop images with class-imbalance due to the small size of disease regions (red: healthy plant / green: disease, black: BG). The left is maize plants with leaf blight, and the right is cabbages with clubroot.

recognize narrow lanes for safety. Similarly, in healthcare, it is important to capture small tumors precisely for diagnosis. Therefore, the segmentation of small RoI classes becomes a critical hurdle that should be overcome in various industrial applications.

To address the issue, we propose the concept of RoI-Attention, which guides the network to intensively reconsider regions with a high likelihood of being predicted as the RoI class (the disease class in this paper) and their surroundings. RoI-Attention is implemented by providing additional supervision to the network for masked version of input image which only represents the regions predicted as disease and their surroundings. This enables the RA-Net to perform double-checking for RoI-associated context. The masked image is not generated through separate preprocessing but by utilizing the initial segmentation result for the input image, which corresponds to probabilistic class estimation. Then, the fusion stage combines the features extracted from the global input image and its masked version, enhancing representation power for the small RoI class. Through a simple convolution operation on the fused features, the final segmentation output is generated with enhanced prediction performance for the RoI class. As shown in Figure 2, compared to conventional segmentation networks that solely perform optimization for the global input, the proposed RA-Net introduces a sequential framework that includes double-checking for RoI-associated context.

We applied the proposed RA-Net for the detection of leaf-blight disease on maize crop images [10]. Experimental results show that the proposed method is effective in improving the predictive performance for the small RoI class of disease.

II. RELATED WORKS

Small Object Segmentation is a challenging task, and it demands a precise identification of class-limited number of pixels in an image and is a lack of geometric cues, resulting in weak feature representation and low 'objectness'. Therefore,

TABLE 1. Notation used in the paper.

Notation	Definition
I	Input RGB image
P_{init}	Initial estimation of class probabilities from I
F_{init}	Feature extracted from I
I_{RoI}	RoI masked version of I using P_{init}
P_{RoI}	Estimation of class probabilities from I_{RoI}
F_{RoI}	Feature extracted from I_{RoI}
P_{fin}	Final segmentation prediction
CE	Cross-Entropy loss

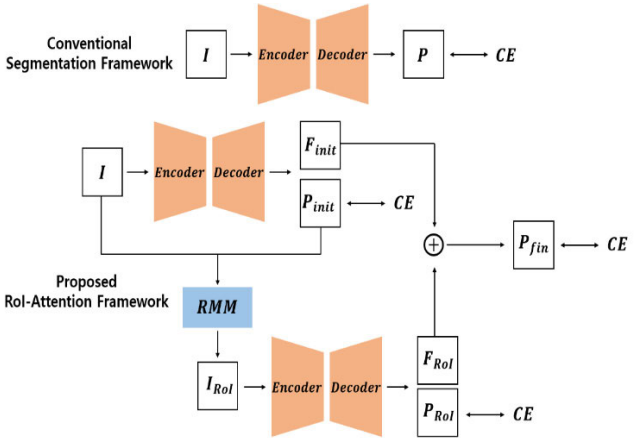

FIGURE 2. Comparison between the conventional segmentation framework and the proposed RoI-Attention framework.

it requires more contextual reasoning for accurate detection. The issue of small object segmentation can be more problematic in the disease segmentation of crop images, which is the primary concern of our study.

Various studies aim to improve the segmentation performance of small objects. BiSeNetV2 [11] tackles the issue of losing representation power for small objects during down-sampling by incorporating the detail-branch. This module enhances the descriptor for pixels by providing higher resolution and deeper channel dimensions. STDC-Seg [12] complements the encoder-decoder structure with a detail-aggregation module, leveraging fine-grained information from edges of objects. HRNet [13] prevents spatial information loss by integrating an additional sub-network which maintains high-resolution features during encoding.

In addition, various methods such as DDRNet [14], MCINet [15], SFNet [16], SPiN [17], MFNet [18] utilize multiscale and high-resolution features or employ additional modules and branches to provide advantageous guidance for recognizing small objects.

While these techniques may be effective for small object segmentation, they have the drawback of not being memory efficient due to the use of additional modules or networks. Compared to these methods, the proposed RA-Net efficiently improves the recognition ability for small objects without relying on additional modules and is even based on a lightweight backbone [19].

Attention-Based Methods have made remarkable advancements in the field of computer vision. References [20], [21], [22], [23], [24], and [25] are studies in medical image processing and they employ Transformer attentions to recognize small organs or cells in CT images more accurately. In the field of agriculture, [8] improved the predictive capability for crops and diseases by using feature refinement with self-attention and feature fusion with cross-attention. These methods aim to overcome the limitations of CNNs, where fine grained information of small objects is lost during the stride or pooling process, by modeling the global relationship among pixels through various attention operations. Furthermore, they enhanced segmentation performances by utilizing multi-scale feature strategy, like UNet [26] or DeepLabV3+ [27].

While those attention-based networks may be useful for accurately detecting small objects, they still pose a challenge in terms of memory efficiency. The Transformer module requires complex computations to consider long-term dependencies among pixels, and this can lead to an intensive increase in the number of model parameters. As the network becomes more sophisticated, the amount of data required for optimization tends to increase significantly. For instance, vision transformer based DPT [28] requires a significant amount of data for effective training. Because making labels for segmentation is labor-intensive, this can become an obstacle from a cost perspective. Reducing the resolution of the input image can alleviate the complexity of the network to some extent but considering that the disease regions are small RoI class, this is not an appropriate solution.

Compared to those existing methods, the proposed RA-Net implements attention based on the network output of the initial input image. It is noteworthy that the RA-Net can be simply implemented by reusing the network's output with class probabilities without complex modules, while remaining faithful to the concept of attention that focuses on important RoI-associated context of the image. Moreover, it possesses advantages in terms of efficiency and performance, compared to transformer networks that prioritize efficiency like Lawin [29], SegNeXt [30], EfficientViT [31].

ROI-Attention aims to guide the network's focus towards more important context within an image. Reference [32] trains the discriminator to determine authenticity only for the RoI (leaf) objects of the given images to generate more plausible leaf images. It masks generated or real images to allow the discriminator to consider leaf objects only, alleviating the effort in considering relatively less important non-leaf regions. In this process, the output of a binary segmentation network is used as a mask to exclusively represent the leaf objects in the input image for the discriminator.

Similarly, the proposed RoI-Attention uses the output of the segmentation network for the global input image to calculate a RoI-Attention map. The estimated map is then used in the subsequent steps to generate an RoI-attentive image as shown in Figure 3 or 6. The difference is that the RoI-attentive image considers not only the RoI predicted regions but also their surroundings to make the RA-Net double-check RoI-associated context.

Reverse-Attention [33] is the opposite concept to RoI-Attention. While RoI-Attention guides the network to focus on regions with high likelihood of being predicted as the RoI class and their surroundings, reverse-attention directs the network to discover unseen regions by erasing the current predicted regions. Note that the RA-Net is trained to reconsider important context in the input image based on the current predicted regions.

In this paper, we experimentally verified that the proposed RoI-Attention, which focuses on RoI-related context, is more superior to reverse-attention in detecting small RoI regions. In addition, RoI-Attention is more versatile as it can be applied to multi-class segmentation (plant, disease, and BG in this paper), making it more useful compared to the reverse-attention based methods [33], [34], [35] that only assume binary segmentation.

III. PROPOSED METHOD

Our task is multi-class segmentation of crop images [10] with three classes (plant, disease, and BG). The most crucial issue is the small size of the RoI class, disease, which makes accurate detection challenging. The goal of our study is to improve the segmentation performance of small disease regions in crop images. The core idea is to guide the network to re-segment the RoI-associated regions which exclusively contain initially predicted regions as the disease class and their surroundings. For this purpose, we generate an RoI-attentive image I_{RoI} that contains the only RoI-associated context.

The RoI-attention map that selectively extracts RoI-related context is produced by RMM (RoI-Mask Module) using the initial segmentation, P_{init} of the global input image, I. The attention map serves as a mask that covers less important context in predicting disease regions from I. Therefore, RoI-Attention can be easily implemented leveraging I_{RoI}, the masked version of I, as an additional input. It is noteworthy that I_{RoI} relies solely on P_{init} which is generated by optimizing the cross-entropy loss without any separate preprocessing.

With the RoI-attentive image I_{RoI} as an additional input, the RA-Net can be trained to reconsider RoI-related context necessary for accurately segmenting disease regions. Therefore I_{RoI} serves as an augmented data since the network, trained on I, implements additional training for I_{RoI}. Then, the feature representations F_{init} and F_{RoI} for the global input I and RoI attentive image I_{RoI} are finally combined to enhance the representation power for disease regions. Utilizing the fused feature, the RA-Net can produce a final output P_{fin} which achieves more accurate performance compared to P_{ori}. We used ERFNet [19] as a baseline network that constituting the RA-Net. In this section, we provide the detailed explanation of the proposed RA-Net architecture in Figure 3 and RMM in Figure 6.

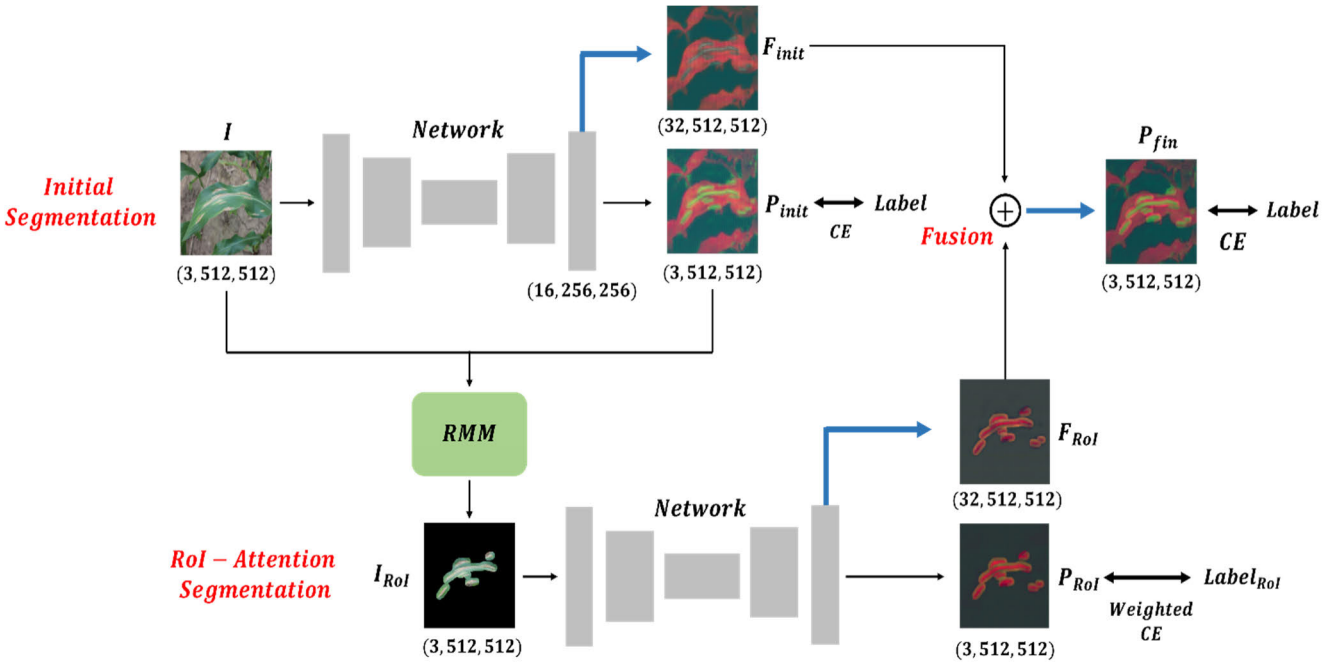

FIGURE 3. The overall architecture of RA-Net. RMM is a RoI-Mask Module which is also depicted in Figure 4. Blue arrow means 1 × 1 convolution, and ERFNet was used for network.

A. RA-NET

RoI-Attention is implemented by using the output of the initial segmentation network P_{init} for the global input image I. Furthermore, to generate a final prediction, it is essential to combine the feature embedding outputs of both the initial and RoI-Attention segmentation stages. Therefore, RA-Net consists of three sequential stages in total: the initial segmentation stage, the RoI-Attention segmentation stage, and the fusion stage. Each stage is optimized for the corresponding GT. This process can be summarized as refining the prediction of the initial segmentation stage by using the RoI-Attention segmentation and fusion stage. Note that the networks used in the first and second stages share weights and differ only in their inputs. The final fusion stage fuses feature embedding outputs F_{init} and F_{RoI} from the stage one and two to improve the feature representation for the RoI-associated context. In the following sections, we will provide detailed descriptions of three sequential stages.

B. INITIAL SEGMENTATION STAGE

This stage produces high-dimensional feature embedding, F_{init} which is used in the fusion stage and the segmentation output, P_{init} corresponding to class probabilities. The feature embedding is from the last decoding layer of the network, as it has not only high resolution but also sufficient semantic information. Note that F_{init} has c channels and it is set to 32. As the stage is identical to the training process of a conventional segmentation network, except for the part that calculates F_{init}, the optimization of P_{init} suffers from the coarse disease prediction. The purpose of our study is to generate P_{fin}, which shows higher predictive performance for disease than P_{init}, using additional stage with RoI-Attention.

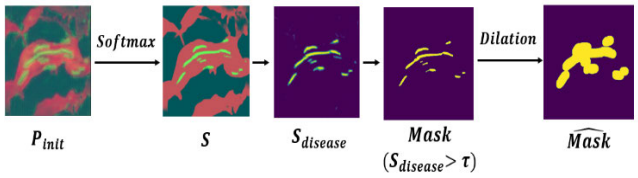

FIGURE 4. Visualized process of RMM. The $\widehat{Mask}$ is used as RoI-Attention map.

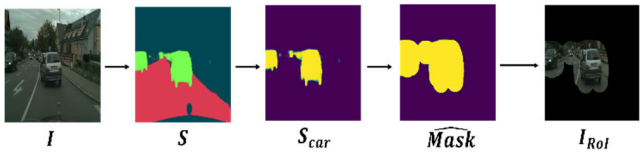

FIGURE 5. Visualized process of RMM for the car object in the Cityscapes dataset.

To implement the RoI-Attention, P_{init} and I are passed to the RMM of the second stage.

C. RoI-ATTENTION SEGMENTATION STAGE

This stage performs the proposed RoI-Attention, that can alleviate the burden on the network by concentrating on the disease-associated context without less important plant and BG information. To calculate the RoI-Attention map which generates I_{RoI} by masking I, P_{init} from the initial stage is used. Table 2 and Figure 4 show the process of RMM to generate the RoI-Attention map. In addition, Figure 5 demonstrates that the proposed RMM is applicable to images on various domains. It simply provides the process of generating I_{RoI} for the segmentation in autonomous driving using the Cistyscapes dataset [43].

TABLE 2. The overall process of RMM.

RoI Mask Module
1. Get class probabilities P_{init} from I
2. Calculate score map $S = Softmax(P_{init})$
3. Extract the disease score map $S_{disease}$ from S $S_{disease} = S[:, disease, :, :]$
4. Get a $mask$ from $S_{disease}$ by thresholding with τ $mask = 1\ if\ S_{disease} > \tau\ and\ 0\ for\ otherwise$
5. Apply dilation operation to the $mask$ with m $RoI\ Attention\ map = \widehat{mask} = dilate(mask\,; m)$
6. Get masked image $I_{RoI} = I \times \widehat{mask}$ and its corresponding GT $label_{RoI} = label \times \widehat{mask}$

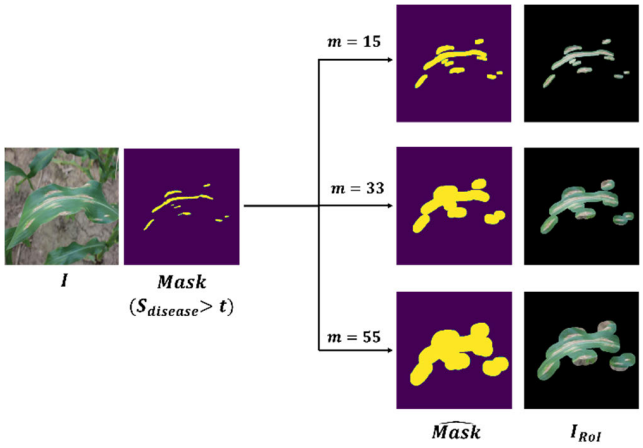

FIGURE 6. Attention map and its I_{RoI} according to the dilation margin m.

RMM first calculates the confidence score map S by applying the *softmax* function to P_{init} along the channel axis. Each channel of S represents the probability of each pixel in the image being predicted as one of the classes: plant, disease, and BG. As our focus is placed on disease-associated context at this stage, we extract the score map corresponding to the disease class only, denoted as $S_{disease}$. The disease score map $S_{disease}$ has values between 0 and 1, as each score represents the probability of being predicted as the disease class. It generates a *mask* by extracting regions with scores higher than a specific threshold τ, thereby containing only the regions with high confidences of the disease class.

Note that, the proposed RoI-Attention stage additionally re-segments regions with high likelihood of being predicted as the disease class but also their surroundings. Therefore, by applying a dilation operation to the *mask*, surroundings of the predicted disease region can be included. A dilated mask $\widehat{mask}$ is used as an RoI-Attention map, and I_{RoI} is calculated by multiplying I with $\widehat{mask}$. The dilation operation of [36] was used and the margin m adjusts the dilation rate. A larger margin allows a broader surrounding context of the disease predicted regions as shown in Figure 6.

Through this process I_{RoI} can be easily calculated from the output of the initial segmentation stage, which is trainable. The network used in the initial stage takes I_{RoI} as an additional input and outputs F_{RoI} and P_{RoI}. Through this additional supervision, the network can double-check disease-associated context, thereby improving the disease recognition ability of the network. Furthermore, it can also enhance its training performance by the augmented data I_{RoI}.

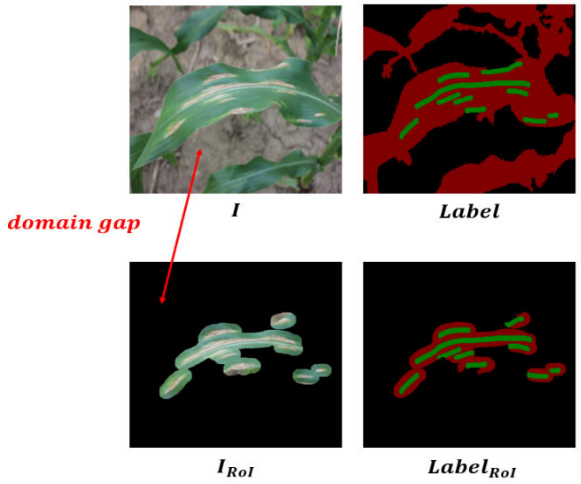

FIGURE 7. Domain gap in BG classes between I and I_{RoI}, (black: BG, red: plant, green: disease).

D. FUSION STAGE

This stage takes F_{init} and F_{RoI} from the first and second stage to produce a final prediction P_{fin}. In fact, this stage complements the shortcomings of the initial segmentation stage, which is a typical training process of a segmentation network. By adding F_{RoI} to F_{init}, the representation power of F_{init} for the disease-associated context can be enhanced. Afterward, a simple convolution operation is performed to produce P_{fin}, which achieves better predictive performance on disease rather than P_{init}.

E. TRAINING

There exist three segmentation outputs $\boldsymbol{P_{init}}$, $\boldsymbol{P_{RoI}}$, and $\boldsymbol{P_{fin}}$, in total and each output is optimized with the cross-entropy loss for its corresponding GT. However, there is a remarkable point to consider when training $\boldsymbol{P_{RoI}}$. Since $\boldsymbol{I_{RoI}}$ is a masked version of $\boldsymbol{I}$, there inevitably occurs a significant loss of meaningful information in the image. As shown in Figures 3 and 5, most of pixel values in $\boldsymbol{I_{RoI}}$ are 0, excluding the disease predicted regions and their surroundings. This can be problematic when calculating the cross-entropy loss for $\boldsymbol{P_{RoI}}$, as $\boldsymbol{I_{RoI}}$ has not only a significant imbalance between the BG (pixels with 0) and the other classes but also a prominent domain gap between the BG classes when compared to $\boldsymbol{I}$. This can hinder the stable convergence of the network. As shown in Figure 7, in $\boldsymbol{I}$, the BG class includes weeds and bare soil, while in $\boldsymbol{I_{RoI}}$, pixels with a value of 0 are treated as the BG class.

To address this issue, when calculating the cross-entropy loss for $\boldsymbol{P_{RoI}}$, we utilize pixel weights, giving more weights to the regions predicted as disease. We use the score map of the disease class ($\boldsymbol{S_{disease}}$ in Table 2 and Figure 4) as pixel

weights, which are calculated from the output of the initial segmentation stage. As $\boldsymbol{I_{RoI}}$ focuses on regions where disease class is likely to be predicted in $\boldsymbol{I}$, assigning higher weights to pixels with high confidence scores for the disease class is a valid approach. This allows for the natural exclusion of unnecessary regions (pixels with 0 in $\boldsymbol{I_{RoI}}$) during the loss calculation. It is possible to reduce calculation errors for the BG class when the cross-entropy loss for $\boldsymbol{P_{RoI}}$ is computed. In summary, the total loss is given by as below.

$$\boldsymbol{L_{total}} = -\frac{1}{\boldsymbol{H}\cdot\boldsymbol{W}}\Big(\sum_{i=1}^{\boldsymbol{H}}\sum_{j=1}^{\boldsymbol{W}}\sum_{c=1}^{\boldsymbol{C}}\Big[\boldsymbol{y_{ij}} - \boldsymbol{log}\left(\boldsymbol{p_{ij}^{init}}\right)\Big] + \boldsymbol{w_{ij}}\cdot\Big[\boldsymbol{y_{ij}^{RoI}} - \boldsymbol{log}\left(\boldsymbol{p_{ij}^{RoI}}\right)\Big] + \Big[\boldsymbol{y_{ij}} - \boldsymbol{log}\left(\boldsymbol{p_{ij}^{fin}}\right)\Big]$$

where $\boldsymbol{y_{ij}}$ is GT, $\boldsymbol{P_{ij}}$ is a network output, and $\boldsymbol{w_{ij}}$ is a pixel weight.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTINGS

The proposed network is trained with the field images infected with leaf blight of the maize plant dataset [10]. There are 110 images for training in total and 10 images for validation. In training, the resolution of the input image is 512 × 512 and batch size 3 are used for training. The proposed RA-Net was designed by extending ERFNet [18] which is known for its lightweight structure. It was implemented using the PyTorch framework on a PC with NVIDIA RTX 3090 GPU. We adopted the Adam optimizer for loss optimization, and initial learning 1e-3.

B. QUANTITATIVE EVALUATION WITH EXISTING METHODS

Various experiments have been conducted to demonstrate that proposed RA-Net is superior for detecting small RoI disease. As an evaluation metric, IoU (Intersection over Union) for each plant and disease class was used. It represents the ratio of the overlapping area between the predicted segmentation result and the GT label to the entire area. As our primary focus is placed on disease regions, and the prediction of the plant regions is relatively easy, we paid more attention to the disease IoU for evaluating the predictive performance of the methods. Furthermore, among the comparison methods utilized, the top 5 with superior disease IoU were compared using different evaluation metrics as shown in Table 6, namely Symmetric Best Dice(SBD) [44] and Dice coefficient(Dice) [45]. Both evaluation metrics indicate that the RA-Net, excels in detecting disease.

First, we show the superiority of our RA-Net by comparing it with the baseline network ERFNet [19], and the networks specialized in small object segmentation. As shown in Table 3, the proposed network demonstrates superior performance in disease segmentation while being efficient in terms of the number of parameters. Without the hierarchical structure [17], [26], multi-scale feature strategy [13], [14], [15], [18], [27], and additional modules or guidance [11], [12], [13], [16], [17], RA-Net effectively improves the representation power of disease. It can also be observed that

TABLE 3. Comparison with baseline (ERFNet) and the existing small object segmentation methods.

Method	Plant IoU	Disease IoU	# parameters
RA-Net	**0.837**	**0.431**	**2,109,026**
ERFNet [19]	0.838	0.366	2,063,151
BiSeNetV2 [11]	0.837	0.416	5,193,263
STDC-Seg [12]	0.840	0.398	4,148,246
HRNet [13]	0.832	0.390	29,534,467
DDRNet [14]	0.832	0386	5,732,262
MCINet [15]	0.843	0.312	27,690,959
SFNet [16]	0.842	0.395	13,915,843
SpiN [17]	0.837	0.399	5,228,369
MFNet [18]	0.844	0.382	1,335,190
UNet [26]	0.842	0.413	15,497,770
DeepLabV3+ [27]	0.830	0.406	40,470,720
FCHardNet [37]	0.846	0.393	4,119,257

RA-Net is more efficient and performs better than [37], which focuses on network lightweighting.

Second, many of Transformer-based methods rely on highly complex operations, leading to heavy-weight architectures. As confirmed in Table 4, when compared with the Transformer based methods which require more parameters than the baseline [19], the proposed network is not only memory efficient but also effectively guides the network to focus on disease-associated context. This demonstrates that in case with limited number of training data, the use of complex operations can hinder the proper convergence of the network as in [41] or fail to deliver the expected improvements. For instance, [29] and [30], despite emphasizing efficiency as a strong point, require more parameters than the RA-Net and exhibit underfitting in terms of predictive performance. In addition, despite using the large number of network parameters, [28] exhibits poor performance, compared to the RA-Net. The cases of [28] and [41] suggest that complex networks with large parameters may not be always helpful for good crop image segmentation, particularly for insufficient amount of training data available. Because masking segmentation labels is generally labor-intensive, complex Transformer-based methods is not appropriate for smart agriculture. In this regard, the proposed RA-Net has the advantage of being data-efficient.

In addition, there also exist Transformer-based approaches [25], [31] with a fewer number of parameters, compared to the baseline. However, they show inferior performance to the RA-Net. This implies that focusing on the efficiency and simplicity of the network can lead to a failure in achieving the most critical goal, improving disease prediction performance. In contrast to these methods, our proposed RA-Net effectively avoids the underfitting in disease prediction.

Third, the proposed RoI-Attention has a concept opposite to Reverse-Attention based methods [33], [34], [35].

TABLE 4. Comparison with Transformer-based methods.

Method	Plant IoU	Disease IoU	# parameters
RA-Net	**0.837**	**0.431**	**2,109,026**
ERFNet [19]	0.838	0.366	2,063,151
Trans-UNet [20]	0.834	0.400	67,865,747
UCTransNet [21]	0.843	0.408	67,225,667
UNeXt [22]	0.843	0.408	67,225,667
TransDeepLab [23]	0.832	0.373	21,153,924
TransAttUNet [24]	0.841	0.372	25,966,028
SAUNet [25]	0.833	0.379	482,870
DPT [28]	0.843	0.396	124,001,326
Lawin [29]	0.813	0.287	18,493,171
SegNeXt [30]	0.823	0.260	27,560,899
EfficientViT [31]	0.841	0.393	704,627
MANet [38]	0.841	0.410	29,610,683
SegFormer [39]	0.834	0.400	67,865,747
GAUNet [40]	0.841	0.410	29,610,683
SETR [41]	0.786	0.221	86,999,81

TABLE 5. Comparison with reverse-attention based methods.

Method	Plant IoU	Disease IoU	# parameters
RA-Net	**0.837**	**0.431**	**2,109,026**
Reverse-Attention [33]	0.837	0.408	24,697,989
PraNet [34]	0.835	0.412	32,547,319
CaraNet [35]	0.841	0.417	46,642,560

TABLE 6. Comparison with To5 methods using SBD & Dice.

Method	SBD	Dice
RA-Net	**0.811**	**0.593**
BiSeNetV2	0.784	0.574
GAUNet	0.778	0.553
MANet	0.773	0.548
UCTransNet	0.776	0.557
UNeXt	0.770	0.539

While Reverse-Attention focuses on unseen regions based on the current prediction, RoI-Attention focuses on the regions predicted as disease and their surroundings in the initial prediction. As shown in Table 5, the proposed RA-Net shows better predictive performance for disease than [33], [34], and [35]. Since they only assume binary segmentation, we performed binary segmentation separately for the plant and disease classes and then aggregated the results.

Lastly, to generate the final prediction $\boldsymbol{P_{fin}}$, RA-Net requires sequential optimization of the outputs from $\boldsymbol{I}$ and $\boldsymbol{I_{RoI}}$ in the preceding two stages. Therefore, ERFNet, the baseline network, consists of RA-Net is trained with both $\boldsymbol{I}$ and $\boldsymbol{I_{RoI}}$. As shown in Table 7, the inference results derived from $\boldsymbol{P_{init}}$ showed superior performance compared to ERFNet which is trained with $\boldsymbol{I}$ only. Therefore, it was confirmed that $\boldsymbol{I_{RoI}}$, which contains the disease-associated context only, can serve as an effective data augmentation technique for training disease segmentation network. Furthermore, it was observed that the proposed RoI-Attention framework is more effective in addressing class-imbalance compared to the OHEM Loss (Online Hard Example Mining) [42].

TABLE 7. Comparison with P_{init} from the initial stage.

Method	Plant IoU	Disease IoU
RA-Net (P_{fin})	0.837	0.431
RA-Net (P_{init})	0.836	0.405
ERFNet [19]	0.838	0.366
ERFNet + OHEM [42]	0.841	0.417

TABLE 8. Comparison with a margin m.

Margin	Plant IoU	Disease IoU
$m = 0$ (no dilation)	0.840	0.408
$m = 15$	0.840	0.422
$m = 33$ (proposal)	**0.837**	**0.431**
$m = 55$	0.840	0.426

TABLE 9. Comparison with channel dimension c.

Channel Dimension	Plant IoU	Disease IoU
$c = 16$	0.836	0.423
$c = 32$ (proposal)	**0.837**	**0.431**
$c = 64$	0.841	0.405

TABLE 10. Comparison with various losses for P_{RoI}.

Loss for P_{RoI}	Plant IoU	Disease IoU
Weighted Cross Entropy(proposal)	**0.837**	**0.431**
Cross Entropy	0.838	0.417
OHEM	0.836	0.423

V. ABLATION STUDIES

There exist two hyper-parameters for the implementation of RA-Net. One is the margin $\boldsymbol{m}$ used in the RoI-Attention stage with RMM which determines the dilation rate. The

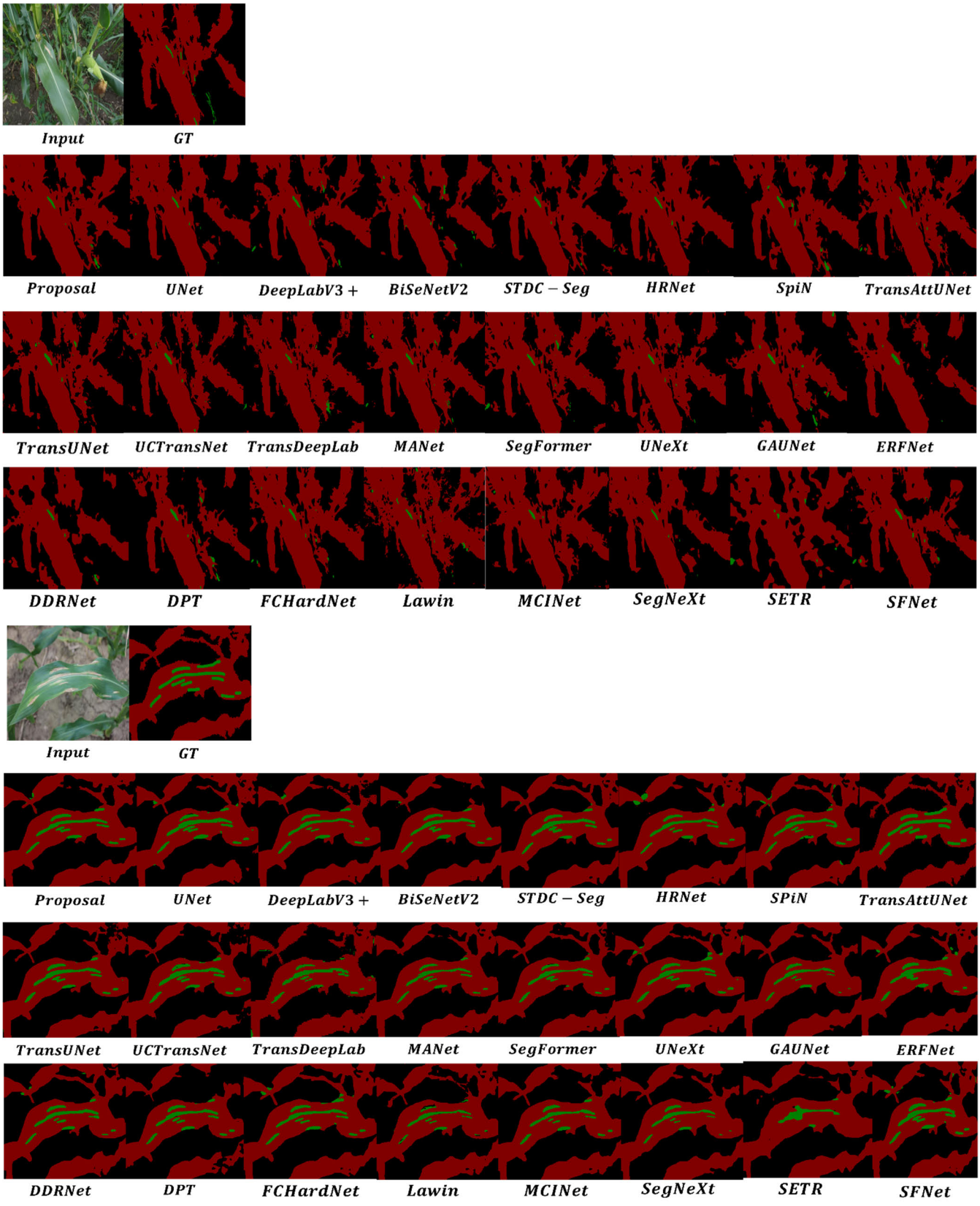

FIGURE 8. Qualitative comparison with the existing methods.

other is the channel dimension c of feature embeddings at the decoding stage. To investigate the effect of hyper-parameters on the performance of the network, ablation studies were conducted.

First, the effect of the margin was verified. As shown in Table 8, employing an RoI-Attention map with margin has significant benefit compared to the case of no margin which considers RoI predicted regions only. This stems from the fact that disease predictions at the initial segmentation stage are coarse during the early stages of training. Therefore, if the only disease predictions are used as the RoI-Attention map, it becomes challenging to obtain meaningful $\boldsymbol{I_{RoI}}$ due to the coarse predictions in the early training stages. By incorporating the surroundings of regions with disease predicted regions through dilation operations, it is possible to address the issues caused by coarse predictions of early training stages. It was also observed that increasing the margin beyond a certain threshold does not significantly contribute to performance improvement.

Second, the channel dimension of feature embedding can be considered as the number of feature descriptors for each pixel. As shown in Table 9, increasing the channel dimension sufficiently can provide a slight improvement in detecting small disease area. However, excessively increasing the channel dimension can lead to overfitting, so it requires careful tuning to find an appropriate balance.

Lastly, when calculating the cross-entropy for the P_{RoI}, the score map of the disease class $S_{disease}$derived from P_{init} is used as pixel weights. The experiments have verified that the weighted cross-entropy for the optimization of P_{RoI} is beneficial for training RA-Net. Table 10 shows the effectiveness of the weighted cross-entropy in addressing class imbalance and the domain gap issue of the BG class which occurs during the RMM process.

VI. CONCLUSION

In this paper we proposed a novel framework to address the challenge of small RoI objects in segmentation. Through our proposed RA-Net, we have confirmed its superior performance in detecting small disease regions compared to existing methods. RoI-Attention, the core idea of RA-Net, guides the network to re-examine the regions predicted as disease and their surroundings. This approach enhances the network's recognition ability for small disease regions without relying on complex modules such as Transformer. Our study is note-worthy in that RoI classes in segmentation, are mostly local regions that are hard to detect. Therefore, our proposed framework is not limited to segmentation for smart agriculture and is easily applicable to any other segmentation tasks.

The current RoI-Attention implemented by the RoI-Mask Module (RMM), utilizes dilation operation that is dependent on hyperparameters. It needs improvement given that the appropriate setting of the dilation margin can have an impact on performance. Future work will be done by exploring methods that can adaptively consider the RoI-related context without relying on hyper-parameters.

Furthermore, although segmentation is a useful method for smart agriculture, there is currently a scarcity of datasets suitable for disease segmentation. And most of them are not from the real field environments but rather from laboratory settings. Our follow-up research aims to contribute to the advancement of deep-learning based smart agriculture.

REFERENCES

[1] J.-Y. Jung, S.-H. Lee, T.-H. Kim, M.-M. Oh, and J.-O. Kim, "Shape based deep estimation of future plant images," *IEEE Access*, vol. 10, pp. 4763–4776, 2022, doi: 10.1109/ACCESS.2022.3140464.

[2] T. Kim, S.-H. Lee, and J.-O. Kim, "A novel shape based plant growth prediction algorithm using deep learning and spatial transformation," *IEEE Access*, vol. 10, pp. 37731–37742, 2022, doi: 10.1109/ACCESS.2022.3165211.

[3] J.-Y. Jung, S.-H. Lee, and J.-O. Kim, "Plant leaf segmentation using knowledge distillation," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2022, pp. 1–3, doi: 10.1109/ICCE-Asia57006.2022.9954844.

[4] S.-E. Lee, S.-H. Lee, and J.-O. Kim, "Deep-clustering based plant disease segmentation network," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Feb. 2023, pp. 1–3, doi: 10.1109/ICEIC57457.2023.10049898.

[5] S.-H. Lee, M.-M. Oh, and J.-O. Kim, "Plant leaf area estimation via image segmentation," in *Proc. 37th Int. Tech. Conf. Circuits/Syst., Comput. Commun. (ITC-CSCC)*, Jul. 2022, pp. 1–3, doi: 10.1109/ITC-CSCC55581.2022.9894907.

[6] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Med. Image Anal.*, vol. 36, pp. 135–146, Feb. 2017, doi: 10.1016/j.media.2016.11.004.

[7] S. Jung, S. Choi, M. Azam Khan, and J. Choo, "Towards lightweight lane detection by optimizing spatial embedding," 2020, *arXiv:2008.08311*.

[8] S.-E. Lee and J.-O. Kim, "Multi-scale attention based plant disease segmentation network," in *Proc. Int. Tech. Conf. Circuits/Syst., Comput., Commun. (ITC-CSCC)*, Jun. 2023, pp. 1–4, doi: 10.1109/ITC-CSCC58803.2023.10212849.

[9] G.-Y. Moon and J.-O. Kim, "Crop & match: RoI cropping and feature matching for segmentation of small objects," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2023, pp. 1–4, doi: 10.1109/ICCE-Asia59966.2023.10326435.

[10] K. Garg, S. Bhugra, and B. Lall, "Automatic quantification of plant disease from field image data using deep learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1964–1971, doi: 10.1109/WACV48630.2021.00201.

[11] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Sep. 2021, doi: 10.1007/s11263-021-01515-2.

[12] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9711–9720, doi: 10.1109/CVPR46437.2021.00959.

[13] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[14] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023, doi: 10.1109/TITS.2022.3228042.

[15] X. Xie, X. Pan, F. Shao, W. Zhang, and J. An, "MCI-Net: Multi-scale context integrated network for liver CT image segmentation," *Comput. Electr. Eng.*, vol. 101, Jul. 2022, Art. no. 108085, doi: 10.1016/j.compeleceng.2022.108085.

[16] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, and Y. Tong, "Semantic flow for fast and accurate scene parsing," 2020, *arXiv:2002.10120*.

[17] A. Wong, A. Chen, Y. Wu, S. Cicek, A. Tiard, B. W. Hong, and S. Soatto, "Small lesion segmentation in brain MRIs with subpixel embedding," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2022, pp. 75–87. [Online]. Available: https://arxiv.org/abs/2109.08791, doi: 10.1007/978-3-031-08999-2_6.

[18] M. Lu, Z. Chen, C. Liu, S. Ma, L. Cai, and H. Qin, "MFNet: Multi-feature fusion network for real-time semantic segmentation in road scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20991–21003, Nov. 2022, doi: 10.1109/TITS.2022.3182311.

[19] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018, doi: 10.1109/TITS.2017.2750080.

[20] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[21] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 2441–2449, doi: 10.1609/aaai.v36i3.20144.

[22] Valanarasu, Jeya Maria Jose, and Vishal M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 23–33.

[23] R. Azad, M. Heidari, M. Shariatnia, E. Khodapanah Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "TransDeepLab: Convolution-free transformer-based DeepLab v3+ for medical image segmentation," 2022, *arXiv:2208.00713*.

[24] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. Wai Kin Kong, "TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation," 2021, *arXiv:2107.05274*.

[25] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "SAUNet: Shape attentive U-Net for interpretable medical image segmentation," 2020, *arXiv:2001.07645*.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[27] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Semantic scene segmentation in unstructured environment with modified DeepLabV3+," *Pattern Recognit. Lett.*, vol. 138, pp. 223–229, Oct. 2020, doi: 10.1016/j.patrec.2020.07.029.

[28] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 12159–12168, doi: 10.1109/ICCV48922.2021.01196.

[29] H. Yan, C. Zhang, and M. Wu, "Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention," 2022, *arXiv:2201.01615*.

[30] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," 2022, *arXiv:2209.08575*.

[31] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17256–17267, doi: 10.1109/ICCV51070.2023.01587.

[32] Q. H. Cap, H. Uga, S. Kagiwada, and H. Iyatomi, "LeafGAN: An effective data augmentation method for practical plant disease diagnosis," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 1258–1267, Apr. 2022, doi: 10.1109/TASE.2020.3041499.

[33] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020, doi: 10.1109/TIP.2020.2965989.

[34] D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "PraNet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020, pp. 263–273. [Online]. Available: https://arxiv.org/pdf/2006.11392, doi: 10.1007/978-3-030-59725-2_26.

[35] A. Lou, S. Guan, H. Ko, and M. H. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," *Proc. SPIE*, vol. 12032, Apr. 2022, Art. no. 120320D, doi: 10.1117/12.2611802.

[36] V. Kulikov, V. Yurchenko, and V. Lempitsky, "Instance segmentation by deep coloring," 2018, *arXiv:1807.10007*.

[37] P. Chao, C.-Y. Kao, Y. Ruan, C.-H. Huang, and Y.-L. Lin, "HarDNet: A low memory traffic network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3551–3560, doi: 10.1109/ICCV.2019.00365.

[38] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020, doi: 10.1109/ACCESS.2020.3025372.

[39] A. Safa, A. Mohamed, B. Issam, and H. Mohamed-Yassine, "SegFormer: Semantic segmentation based tranformers for corrosion detection," in *Proc. Int. Conf. Netw. Adv. Syst. (ICNAS)*, Oct. 2023, pp. 1–6, doi: 10.1109/ICNAS59892.2023.10330461.

[40] J. Maria Jose Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," 2021, *arXiv:2102.10662*.

[41] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886, doi: 10.1109/CVPR46437.2021.00681.

[42] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769, doi: 10.1109/CVPR.2016.89.

[43] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223, doi: 10.1109/CVPR.2016.350.

[44] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: Theory and practice when evaluating with dice score or Jaccard index," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3679–3690, Nov. 2020, doi: 10.1109/TMI.2020.3002417.

[45] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, "Continuous dice coefficient: A method for evaluating probabilistic segmentations," 2019, *arXiv:1906.11031*.

GOO-YOUNG MOON received the B.S. degree from the Department of Statistics, Korea University, Seoul, South Korea, in 2021, where he is currently pursuing the M.S. degree in electrical engineering. His current research interest includes instance and semantic segmentation.

JONG-OK KIM (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Korea University, Seoul, South Korea, in 1994 and 2000, respectively, and the Ph.D. degree in information networking from Osaka University, Osaka, Japan, in 2006. From 1995 to 1998, he was an Officer with Korea Air Force. From 2000 to 2003, he was with SK Telecom Research and Development Center and Mcubeworks Inc., South Korea, where he was involved in research and development on mobile multimedia systems. From 2006 to 2009, he was a Researcher with the Advanced Telecommunication Research Institute International (ATR), Kyoto, Japan. He joined Korea University, in 2009, where he is currently a Professor. His current research interests include image processing, computer vision, and intelligent media systems. He was a recipient of the Japanese Government Scholarship, from 2003 to 2006.