Multiple kernel-enhanced encoder for effective herbarium image segmentation

Sanghyuck Lee,[1] Hyeonji Moon,[2] Sangtae Kim,[2] and Jaesung Lee[1,✉]

[1] *Department of Artificial Intelligence, Chung-Ang University, Dongjak-Gu, Seoul, Republic of Korea*

[2] *Department of Biology, Sungshin Women's University, Gangbuk-Gu, Seoul, Republic of Korea*

✉ E-mail: curseor@cau.ac.kr

The neural network proposed here specializes in herbarium image segmentation. The encoder of the proposed model contains multiple kernels of different sizes to address the complex structures of plant components, such as tangled roots and stems. By employing multiple kernel sizes, the convolution block enables multiscale learning, which is underexplored in previous approaches. This design effectively extracts and fuses local and global features, enabling both broad and narrow perspectives on complex structures within herbarium images and thereby improves segmentation performance. The experimental results demonstrate that the proposed model outperforms three conventional models. The source code can be accessed at https://github.com/tkdgur658/herbarim_segmentation_network

Introduction: Plant identification is the process of naming a plant based on taxonomic knowledge and is essential in several fields, such as ecology, pharmacy, and agriculture [1]. In addition to the plant body, herbarium sheets include non-plant components, such as labels with collection information, rulers, and colour palettes [2]. Formal climate change monitoring can be achieved by exploring the information provided by herbarium images; for example, widely distributed Viola species can be used for this purpose [3]. However, the development of automatic plant identification systems, necessitated by the declining number of taxonomists, requires preprocessing to remove non-plant components from plant images, as these elements interfere with model training [4]. Despite the rapid development of neural network-based segmentation, conventional studies on herbarium image segmentation have not explored recent achievements, such as multiscale convolution. Here, we propose a simple but effective neural network for herbarium image segmentation.

Related work: The field of plant segmentation, including herbarium image segmentation, remains at an early stage compared to major areas such as medical image segmentation. Slight modifications to popular baselines such as UNet [5], SegNet [6], and DeepLabV3+ [7] have been widely explored for plant segmentation. UNet was improved by adding three convolutional layers to the decoding blocks, which improves the extraction of characteristics of the diseased leaf [5]. For SegNet-based models, the SegNet pooling indices were combined with a UNet decoder to improve the segmentation of small targets, such as lesions in tomato leaf disease [6]. In addition, SegNet integrated with an attention module to enhance feature emphasis [8]. Finally, DeepLabv3+-based models enhance segmentation by addressing challenges such as uneven lighting and overlapping plant structures [7].

Within plant segmentation, herbarium images have received even less attention than other subfields. UNet has been included in herbarium segmentation experiments, where the training was conducted on 320 images with pre-trained encoder [9]. Meanwhile, DeepLabV3+ was employed for leaf segmentation in herbarium images and compared to detection-based methods for leaf feature extraction [10]. In addition, DeepLabV3+ was evaluated after training on 2685 specimens from 20 different families, encompassing varying resolution, quality, and layout [11]. Finally, four encoder–decoders were compared with a novel network based on visual geometry group [12], demonstrating the superiority of combinations of average pooling and tanh.

According to our brief review, herbarium image segmentation still requires further research to bridge the gap with major areas of semantic segmentation. Multiscale convolution remains underexplored in plant segmentation and has significant potential in segmentation of complex structures, such as those found in herbarium specimens [13].

Proposed method: Unlike natural plant images captured by non-experts, herbarium images have complex objects, such as tangled roots and stems. To address this issue, we designed the proposed model, characterized by convolution layers with multiple kernel sizes. The proposed model comprises an encoding phase E and a decoding phase D, each consisting of five encoding stages $E^1, \ldots, E^5$ and four decoding stages $D^1, \ldots, D^4$. Let the input image be represented by $x \in \{0, \ldots, 255\}^{H \times W \times 3}$, where H and W denote the height and width of the input image, respectively. The first encoding stage involves two 3×3 convolution layers, one 1×1 layer, and a max-pooling layer. Each convolution layer can be followed by a batch normalization layer and rectified linear unit function. The output of the initial encoding stage E^1 is represented by $z^1 = P^1(F^1(x))$, where F^1 denotes the combination of three convolution layers and P^1 denotes the max-pooling operation. The remaining four encoding stages, denoted by $E^2, \ldots, E^4$, each comprise one proposed block, characterized by multiple kernel sizes within a convolution layer, and one max-pooling layer. The proposed block captures multiscale features to address both broad and narrow target aspects, enhancing the segmentation of complex objects. For each encoding stage $i = 2, \ldots, 4$, the output can be formulated as $z^i = P^i(F^i(z^{i-1}))$, where z^{i-1} is the output of the previous stage, and $F^2, \ldots, F^4$ denote the corresponding blocks for the second through fourth stages. Finally, the output of the encoding phase is defined as $z^5 = F^5(z^4) = E^5(z^4) = E(x)$, where F^5 and E^5 refer to the proposed block.

The proposed block improves the complex object segmentation by extracting multiscale features and fusing them to classify each pixel using both broad and narrow perspectives. Employing multiple kernel sizes naturally supports the capture of both local and global features for enhanced segmentation. Before this process, the proposed block transforms the feature maps into a suitable form for multiple kernel sizes through a composition of 1×1 and 3×3 convolutions. The max-pooling operation in the previous stage with fixed computations for small pool sizes may leave unnecessary information from a global perspective. The 1×1 and 3×3 convolutions compress the essential information along the channel and spatial axes, respectively, effectively preparing a feature map for convolution with multiple kernel sizes. The feature map is then divided into four groups, each containing $c/4$ channels, yielding four distinct feature maps. The four feature maps are passed through a convolution layer utilizing kernel sizes of 1×1, 3×3, 5×5, and 7×7, with padding implemented to preserve the original dimensions. The four feature maps are subsequently aggregated to generate a single feature map involving $c/4$ channels by addition. Multiscale features are then fused to extract high-level representations, effectively capturing complex patterns across different scales. Finally, the input feature map is combined with the expanded feature map using a residual architecture with 1×1 convolution to enhance the representation of the original feature scale.

In the decoding phase, the four decoding stages, denoted as $D^1, \ldots, D^4$, aim to reconstruct the high-resolution output. The four stages in the decoding phase include one unpooling layer and a varying number of 3×3 convolution layers. The output of each decoding stage is calculated as $\hat{z}^j = G^j([F^{(5-j+1)}(z^{(5-j)}), U^j(\hat{z}^{j-1})])$, where U^j is the jth unpooling operation, G^j is the composition of convolution blocks in the decoding stage j, $[\cdot]$ represents the concatenation operation, and $\hat{z}^0 = z^5$. The first and second decoding stages contain three 3×3 convolution layers, whereas the third and fourth decoding stages include one and two 3×3 convolution layers, respectively. Each unpooling layer uses pooling indices from the corresponding max-pooling layer during the encoding stage. The feature maps upsampled by the unpooling layer are concatenated with the feature maps from the corresponding encoding stage before max pooling. The concatenated feature maps are then passed through the convolution layers and fed into the next decoding stage. Finally, the segmentation mask is derived by applying a sigmoid activation function to the output and classifying pixels with significance values over a predefined threshold as belonging to the target region. A schematic of the proposed model is presented in Figure 1.

Table 1. The results of the main experiments, with ▼ and ▼▼ denoting that the proposed model passed the paired t-tests at the 0.1 and 0.05 significance levels, respectively. The values in parentheses indicate the average ranking

Model name	mIoU	DC	Precision	Recall
Proposed	**0.847 ± 0.008 (1.20)**	**0.908 ± 0.006 (1.25)**	**0.885 ± 0.013 (1.40)**	0.960 ± 0.005 (1.70)
Shoabib et al. [5]	0.838 ± 0.003 (2.00)▼▼	0.902 ± 0.002 (1.95)▼▼	0.876 ± 0.008 (2.05)	**0.961 ± 0.005 (1.60)**
Kaur et al. [6]	0.829 ± 0.003 (2.80)▼▼	0.898 ± 0.002 (2.80)▼▼	0.872 ± 0.008 (2.55)▼▼	0.954 ± 0.007 (2.70)▼
Hussein et al. [10]	0.765 ± 0.003 (4.00)▼▼	0.859 ± 0.002 (4.00)▼▼	0.836 ± 0.007 (4.00)▼▼	0.910 ± 0.005 (4.00)▼▼

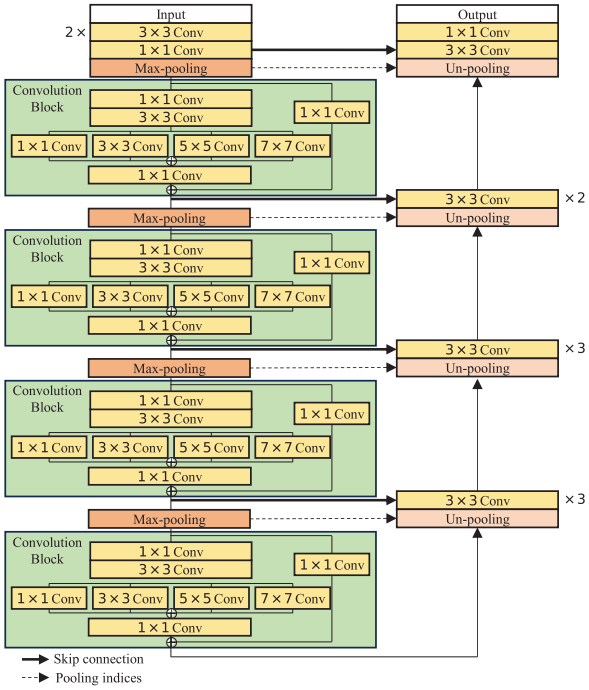

Fig. 1 *Architecture of the proposed model*

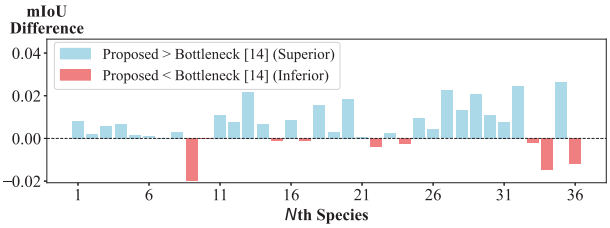

Fig. 2 *Comparison of species-specific mean intersection over union (mIoU) in ablation study. The x-axis were sorted in descending order using mIoU values of the proposed model*

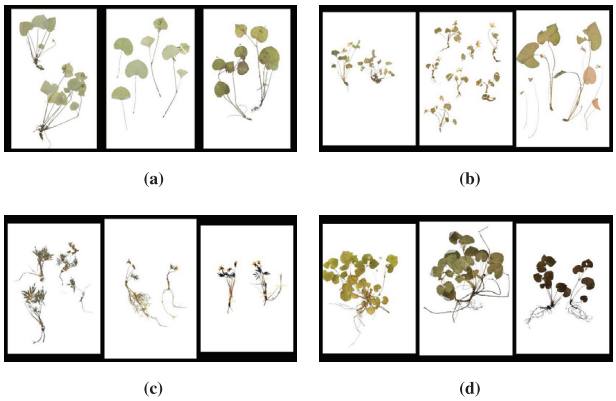

Fig. 3 *Examples from the top two species and the two species showing the most significant improvements. The values in parentheses show the mean intersection over union*

Experimental results: We conducted experiments using three conventional models for plant segmentation [5, 6, 10]. We collected 14,939 specimen images of 36 Viola classes on the Korean Peninsula and conducted 10 iterative experiments by splitting the dataset into training, validation, and test sets in a 0.6:0.2:0.2 ratio. The network input consisted of three-channel RGB images resized to 384 × 256 with zero padding to maintain the original aspect ratio. Each training was performed for up to 50 epochs with an early stopping strategy that halted training if no improvement was observed for 25 epochs. We utilized a batch size of 16 with an AdamW optimizer configured with a learning rate of 1×10^{-3}, and weight decay of 1×10^{-4}. A cosine annealing scheduler was used to adjust the learning rate during training, with a minimum rate of 1×10^{-6}. Dice–cross-entropy loss was used as the loss function without any weighting between the Dice and cross-entropy terms. Data augmentation included random crop resizing with scales from 0.7 to 1.0, horizontal flipping, and random rotations, each applied with 50% probability. We evaluated the models using four common segmentation measures – mean Intersection over Union (mIoU), Dice coefficient (DC), precision, and recall – and performed paired *t*-tests at significance levels of 0.1 and 0.05.

Table 1 represents the comparison results of the proposed and the three comparison models in mIoU, DC, precision, and recall. The experimental results showed that the proposed model significantly outperformed comparison models in mIoU, DC, and precision.

We conducted an ablation study to validate the efficacy of our strategy. Specifically, two counterpart model were used as follows. The first removed the convolution layers with multiple kernel sizes from the proposed blocks (i.e., original bottlenecks [14]), while the second replaced the proposed blocks with UNet blocks, which are widely used in plant segmentation. As shown in Table 2, the proposed model achieved higher values in mIoU, DC, and precision. Furthermore, the robustness of the proposed model is highlighted in Figure 2 with consistent performance improvements in 27 out of 36 species, underscoring the adaptability of the model to diverse species. Notably, these improvements were more pronounced for plants with complex regions. Figure 3a,b shows the random samples for the top two species based on the mIoU values of the proposed and two counterpart models, which were relatively easier targets. Figure 3c,d presents the random samples for two species with the most significant improvements by the proposed block, representing more challenging cases for counterpart models. The targets with greater performance improvement tend to contain more intricate regions

Table 2. The ablation study results, with ▼ and ▼▼ denoting that the proposed model passed the paired t-tests at the 0.1 and 0.05 significance levels, respectively. The values in parentheses indicate the average ranking

Model name	mIoU	DC	Precision	Recall
Proposed	**0.847 ± 0.008 (1.20)**	**0.908 ± 0.006 (1.30)**	**0.885 ± 0.013 (1.55)**	0.960 ± 0.005 (1.75)
Bottleneck [14]	0.841 ± 0.004 (2.20)▼▼	0.904 ± 0.003 (2.20)▼	0.880 ± 0.008 (2.10)	0.959 ± 0.004 (2.35)
UNet Block [15]	0.839 ± 0.004 (2.60)▼▼	0.903 ± 0.003 (2.50)▼▼	0.877 ± 0.008 (2.35)	**0.961 ± 0.005 (1.90)**

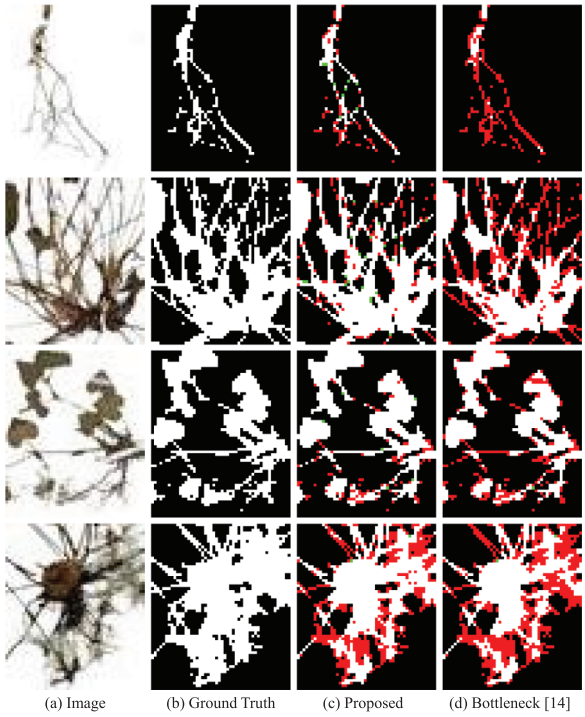

Fig. 4 *Comparison of output masks from the proposed and counterpart models in the ablation study. White, black, red, and green pixels represent true positives, true negatives, false negatives, and false positives, respectively*

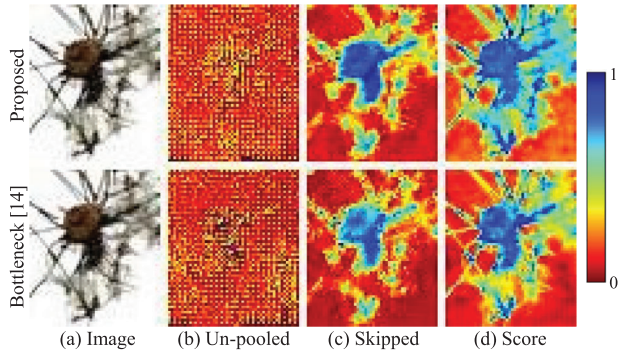

Fig. 5 *The skipped feature maps, un-pooled feature maps, and final score maps for the last decoding stage of the proposed model and its ablation counterpart*

(Figure 3c,d), such as entangled roots or stems, whereas the simpler targets with smaller gains tend to have larger leaves or less complex objects (Figure 3a,b).

Figure 4 illustrates the qualitative results of the ablation study, comparing the output masks of the proposed model and the counterpart model with bottlenecks [14]. The results illustrated an improved performance of the proposed model in recognizing complex objects, such as tangled roots or stems. Finally, Figure 5 illustrates the skipped feature map, un-pooled feature maps fed into the final decoding stage, and the output score maps. The proposed model, with convolutions of diverse kernel sizes, effectively captures fine-grained structures, as evident in the skipped feature map. In contrast, counterpart models struggle to capture such details, resulting in lower-quality output scores.

Conclusion: Here, we propose a novel neural network for herbarium image segmentation. The proposed model outperformed conventional models in the three measures. In future studies, we plan to apply a neural architecture search to optimize the model for this task.

Author contributions: **Sanghyuck Lee**: Software; visualization; writing—original draft. **Hyeonji Moon**: Data curation; formal analysis; software; visualization. **Sangtae Kim**: Conceptualization; formal analysis; funding acquisition; project administration; resources; supervision; validation; writing—review and editing. **Jaesung Lee**: Conceptualization; funding acquisition; methodology; project administration; supervision; validation; writing—review and editing.

Acknowledgements: Due to the characteristics of this study, Sanghyuck Lee and Hyeonji Moon contributed equally as the first authors. In addition, Sangtae Kim and Jaesung Lee contributed equally as the corresponding authors. This work was supported by grants from the National Institute of Biological Resources (NIBRE202411) to Prof. Sangtae Kim and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT, South Korea (2021-0-01341, Artificial Intelligence Graduate School Program [Chung-Ang University]) to Prof. Jaesung Lee.

Conflict of interest statement: The authors declare no conflicts of interest.

Data availability statement: The data used in this study are available from the corresponding author upon reasonable request.

Received: *5 November 2024* Accepted: *20 January 2025*
doi: 10.1049/ell2.70155

References

1 Simpson, M.G.: *Plant systematics*. 3rd ed. Academic Press, Amsterdam (2019)
2 Tan, K.C., et al.: The herbarium challenge 2019 dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–4, IEEE (2019)
3 Marcussen, T., et al.: A revised phylogenetic classification for viola (Violaceae). *Plants* **11**(17), 2224 (2022)
4 de Lutio, R., et al.: The herbarium 2021 half-earth challenge dataset and machine learning competition. *Front. Plant Sci.* **12**, 787127 (2022)
5 Shoaib, M., et al.: Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. *Front. Plant Sci.* **13**, 1031748 (2022)
6 Kaur, P., et al.: Performance analysis of segmentation models to detect leaf diseases in tomato plant. *Multimed. Tools Appl*. **83**(6), 16019–16043 (2023)
7 Polly, R., Devi, E.A.: Semantic segmentation for plant leaf disease classification and damage detection: A deep learning approach. *Smart Agric. Technol.* **9**, 100526 (2024)
8 Cao, L., et al.: Semantic segmentation of plant leaves based on generative adversarial network and attention mechanism. *IEEE Access* **10**, 76310–76317 (2022)
9 White, A.E., et al.: Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. *Appl. Plant Sci*. **8**(6), e11352 (2020)
10 Hussein, B.R., et al.: Automated extraction of phenotypic leaf traits of individual intact herbarium leaves from herbarium specimen images using deep learning based semantic segmentation. *Sensors* **21**(13), 4549 (2021)
11 Weaver, W.N., Ng, J., Laport, R.G.: LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Appl. Plant Sci.* **8**(6), e11367 (2020)
12 Triki, A., et al.: Deep learning based approach for digitized herbarium specimen segmentation. *Multimed. Tools Appl.* **81**(20), 28689–28707 (2022). https://doi.org/10.1007/s11042-022-12935-8
13 Li, J., et al.: MAGF-Net: A multiscale attention-guided fusion network for retinal vessel segmentation. *Measurement* **206**, 112316 (2023)
14 He, K., et al.: Deep residual learning for image recognition. In: CVPR Organizing Committee (eds.) Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, IEEE (2016)
15 Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (eds.) Proceedings of the 18th international conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, Springer, Cham (2015)