

Image Segmentation of Plant Leaves in Natural Environments Based on LinkNet

Lili Zhang^{1, 2, a}, Xiyin Liang^{1, 2, b}

¹ Department of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China

² Gansu Province Intelligent Information Technology and Application Engineering Research Center, Lanzhou, Gansu 730070, China

^a 2796582400@qq.com, ^b silver@nwnu.edu.cn

Abstract: The conventional approach to plant leaf image segmentation for subsequent measurement of leaf geometric parameters, while reasonably accurate, exhibits lower efficiency. To address this challenge, a plant leaf image segmentation algorithm based on deep learning semantic segmentation models and transfer learning is proposed to achieve rapid and precise leaf segmentation. The presented method adopts LinkNet as its foundational network structure and introduces four key enhancements: leveraging ResNet18 as the backbone encoder network to expedite model fitting through transfer learning; reducing the number of encoding and decoding blocks to decrease network complexity; refining channel reduction strategies to minimize parameters in upsampling; and implementing subpixel convolution for upsampling to lower computational load. Integrated with the focal loss function, this improved LinkNet network is applied to a standard leaf dataset. Experimental outcomes indicate that the proposed algorithm achieves a segmentation accuracy of 97.27%, comparable to the original LinkNet; inference time is merely 7.82 ms, a 39.89% reduction compared to the original network; a substantial decrease in model parameters and floating-point operations; and a significantly faster inference speed of the improved network compared to FCN, U-Net, DeepLabV3+, and similar networks. This algorithm not only swiftly segments the main body of leaves but also effectively preserves details such as leaf edge serrations. It truly enables efficient and precise segmentation of plant leaves, offering a novel approach for rapidly measuring leaf area and other geometric parameters.

Keywords: Deep learning; Leaf segmentation; LinkNet; Sub-pixel convolution.

1. Introduction

Leaves are the fundamental components for photosynthesis in plants, serving as the primary organs for synthesizing organic materials [1]. They also represent one of the largest surface areas in contact with the environment. Methods of leaf segmentation based on digital image processing techniques have gained widespread application. For instance, Zhang [2] segmented apple leaf images using color space transformation and the Otsu method. Compared to traditional methods, deep learning exhibits significant transfer learning capabilities and finds extensive applications.

In recent years, with the introduction of classic semantic segmentation models such as Fully Convolutional Network (FCN) [3], U-Net [4] and so on, more scholars are exploring their implementation in agricultural domains. For example, Zhao [5] employed FCN for segmenting diseased grape leaves, achieving automatic leaf image segmentation under complex lighting conditions and backgrounds. Tassis et al. [6] utilized U-Net to segment diseased coffee tree leaves, demonstrating good segmentation results with high accuracy. Hussein [7] applied DeepLabV3+ to segment a comprehensive dataset of plant leaf specimens. In comparison to object detection methods (including Faster R-CNN), the use of deep learning semantic segmentation models yielded superior segmentation results.

In summary, traditional methods exhibit lower efficiency, digital image processing techniques involve complex processes, and semantic segmentation networks often produce subpar segmentation results in detail and require post-processing.

2. Data and Methods

2.1. Dataset

The dataset used in this experiment comprises 174 common classes of apple tree leaves, with each class containing 20 to 50 leaf images, totaling 8,184 images. The images in the dataset have slightly different resolutions, namely 3,216 pixel $\times$ 2,316 pixel and 4,288 pixel $\times$ 2,848 pixel. Training on high-resolution images demands higher device specifications and can slow down network inference speed. Conversely, while low-resolution images expedite both training and inference, they also lead to reduced segmentation accuracy. Hence, for this experiment, the image resolution was standardized to 1,184 pixel $\times$ 800 pixel.

Figure 1-a displays a representative standard apple leaf image. From Figure 1, it can be observed that there's relatively low color contrast between the leaf and the blue background ruler plate; the leaf exhibits numerous and small serrations. Additionally, a small number of leaves in the dataset present natural curling, which collectively adds complexity to the leaf segmentation task.

The foundation for extracting image features using deep learning methods relies on annotating the images, which significantly impacts the segmentation results. In order to preserve the fine details of leaves, such as serrations, while also streamlining the annotation process, this study employed Photoshop for annotation. As depicted in Figure 1-b, leaves (including the petiole) were annotated in green, while the background was annotated in red. As shown in the enlarged portion at the top-left corner of Figure 1-b, the pixel values at the edges of the leaf in JPEG-format annotated images may undergo alterations during transmission and storage,

negatively impacting the segmentation outcome. Hence, prior to experimentation, the annotated images were binarized, as

illustrated in Figure 1-c, and the processed results were saved in PNG format.

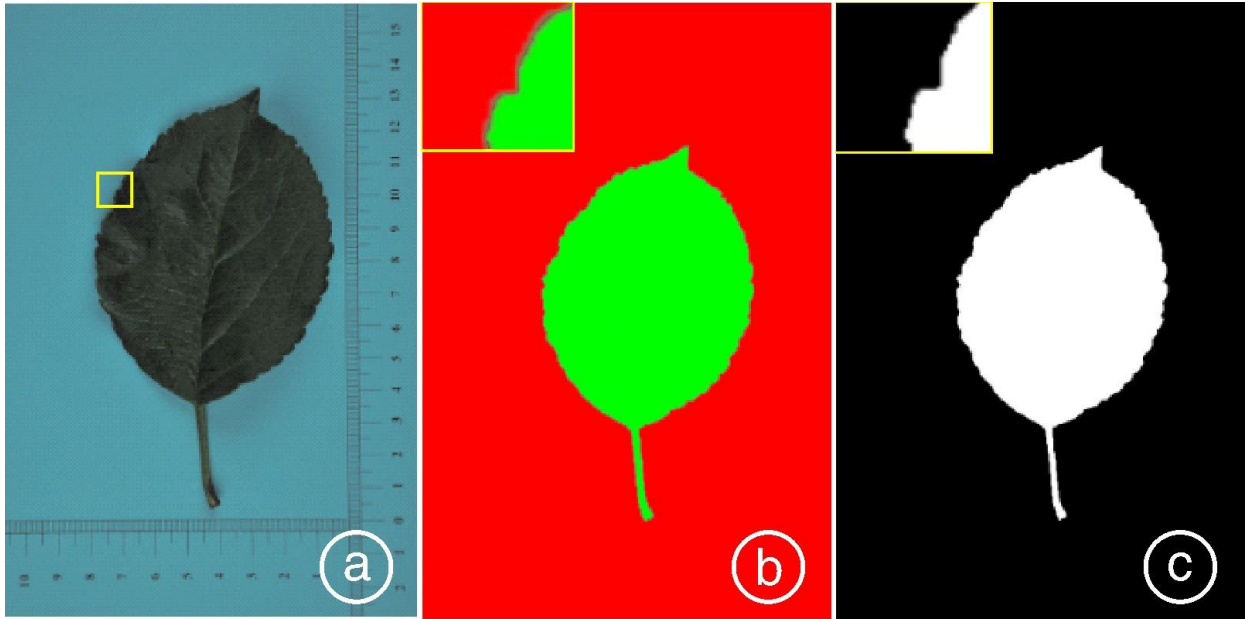


Fig. 1 Data annotation and processing

a, Image; b, Label; c, Binarization. The upper left corners of b and c correspond to the contents of the yellow rectangular box in a, respectively. There was a clear transition line at the junction of the leaf and the background in figure b, which was not present in figure c.

2.2. Training Model

The LinkNet network comprises an encoder network, a decoder network, and a pixel fusion pathway. The encoder network consists of one initial block and four consecutive encoding blocks, implementing downsampling through convolution and max-pooling with a stride of 2 to extract high-level semantic features. Conversely, the decoder network includes four consecutive decoding blocks and one final block, restoring image dimensions through deconvolution to produce segmentation outputs. The pixel fusion pathway denotes the merging of shallow features from the encoder and higher-level semantic features from the decoder via pixel-wise addition. The specific network structure is illustrated in Figure 2.

The initial block, as depicted in Figure 2-b, performs downsampling using a 7×7 convolution and max spatial pooling to extract image features. The internal structure of the encoding block, illustrated in Figure 2-c, comprises two consecutive residual blocks. The first residual block involves a convolution with a stride of 2, reducing the size of the feature map by half while doubling the number of channels. As shown in Figure 2-e, the decoding block encompasses three sequential operations: convolution, deconvolution, and another convolution. The first 1×1 convolution reduces the dimensionality of the input feature map to $1/4$, known as channel reduction, aiming to decrease parameters in upsampling. Subsequently, the feature map is enlarged through deconvolution, followed by the second 1×1 convolution to restore the number of channels, ensuring the smooth integration of output feature maps with global information from the corresponding encoding blocks. Finally, the final block utilizes two deconvolutions to restore the feature map to the original image size and produce the segmentation output.

Due to the large scale and standardization of the leaf dataset used in the experiment, the segmentation results using the original LinkNet network are already quite clear. However,

the model size remains relatively large, and the inference speed is slow, which cannot meet practical agricultural applications. Therefore, this paper has made improvements to the original LinkNet.

Compared to the original LinkNet network, this paper introduces improvements in four aspects: firstly, utilizing the pre-trained model ResNet18 as the backbone encoder to shorten the training time; secondly, removing a corresponding pair of encoding and decoding blocks to reduce the network's depth; thirdly, refining the channel reduction scheme within the decoding blocks to compress the channel count of the pre-upsampling feature map to $1/16$; finally, employing subpixel convolution to replace deconvolution upsampling in the final block. The modified structure of the LinkNet network is illustrated in Figure 3.

The front-end portion of pre-trained deep neural network models, comprising convolutional layers, pooling layers, activation functions, and more, serves the function of feature extraction, aligning with the role of an encoder. It already possesses well-initialized parameter values, making it a suitable choice as the backbone encoder. Utilizing this segment not only addresses the time-consuming nature of training models from scratch but also accelerates model fitting, enabling the network to achieve sufficient training in fewer iterations. Therefore, this paper adopts ResNet18, a model with fewer parameters and higher accuracy, as the backbone encoder, leveraging its pre-trained parameters from ImageNet for transfer learning. This approach expedites model fitting, saving time and computational resources [8].

Although larger network depths and widths can enhance model accuracy, they significantly increase computational requirements, making them less practical for real-world applications [9]. Therefore, to reduce model size and enhance network inference speed, this paper compresses the original LinkNet network. As illustrated in Figure 3, this study reduces the logarithm of LinkNet's decoding-encoding blocks from the original 4 pairs to 3 pairs. Consequently, the network's depth decreases from 33 layers to 25 layers, with a maximum

feature map channel count of only 256. Compared to the original LinkNet, the improved LinkNet not only reduces network depth but also maintains a smaller network width, significantly decreasing network complexity.

Furthermore, in order to reduce the parameter count in the upsampling process and further enhance the network's inference speed, this paper improved the channel reduction scheme within the decoder, as illustrated in Figure 4. Within the improved decoding block, the feature map initially undergoes a 1x1 convolution to reduce the channel count to

1/16 of its original value. Subsequently, it goes through deconvolution upsampling, followed by another 1x1 convolution to restore the channel count for feature fusion. This method reduces the parameter count in the upsampling process without compromising segmentation accuracy. In contrast to the original network, which only reduced the channel count to 1/4 of its original value, the improved channel reduction scheme further decreases parameters in the upsampling process to 1/16 of the original, thereby further enhancing the model's inference speed.

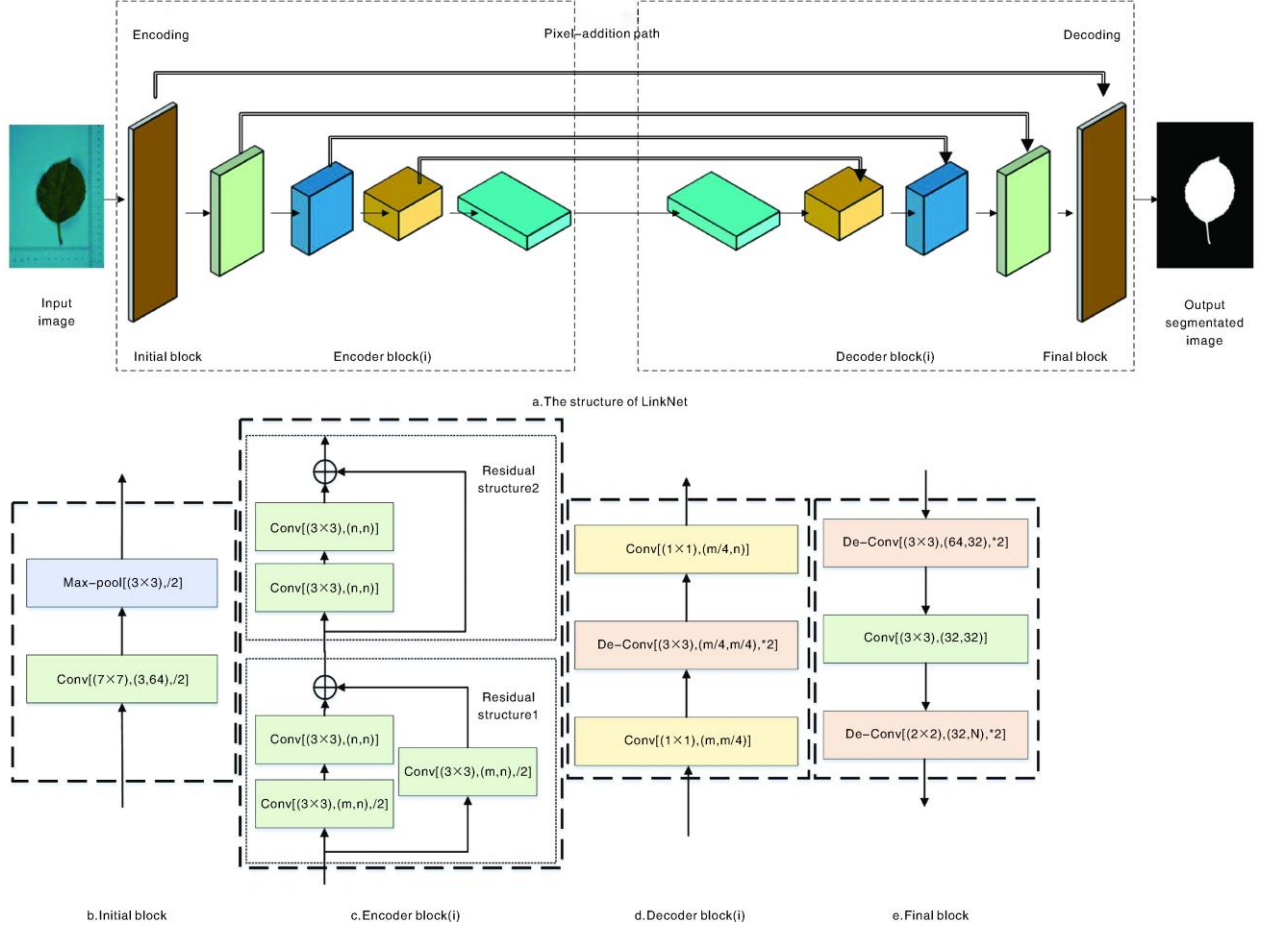


Fig. 2 Framework of LinkNet

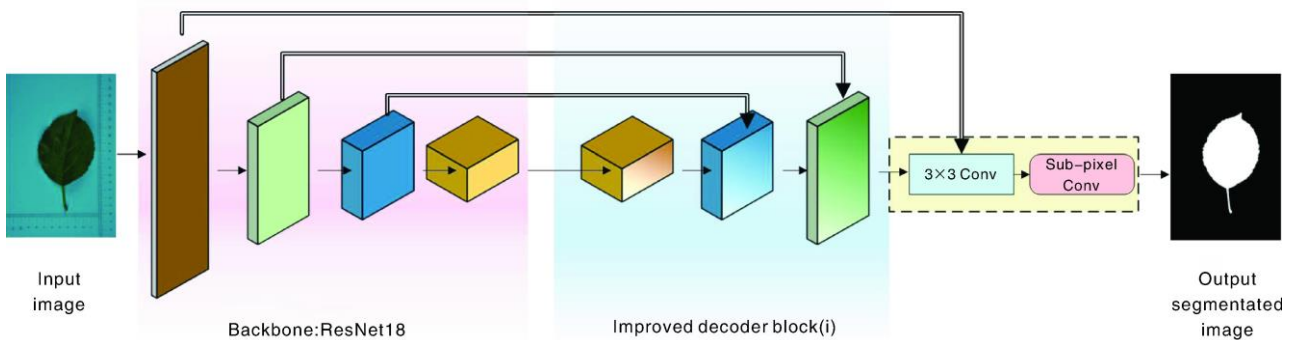


Fig. 3 Framework of improved LinkNet

Finally, the ultimate block in the original LinkNet was replaced with a faster subpixel convolution for upsampling. Subpixel convolution, abbreviated as Sub-pixel conv hereafter, is a commonly used upsampling technique in image super-resolution reconstruction [10].

The principle of subpixel convolution is illustrated in Figure 5. This technique involves rearranging individual

pixels from a low-resolution image into a specific pattern to form a unit in the high-resolution feature map. Consequently, the pixels from the low-resolution image correspond to subpixels in the new feature map, thus termed subpixel convolution. Additionally, the figure depicts that there exists a proportional relationship between the channel counts of the feature maps before and after subpixel convolution: the

channel count of the feature map before upsampling should be r^2 times the channel count after upsampling.

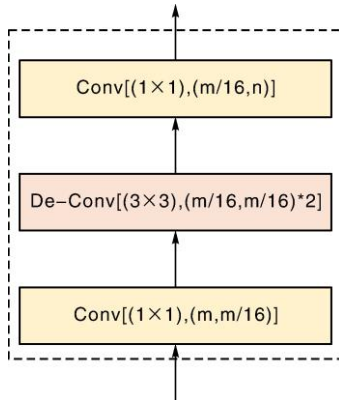


Fig. 4 Improved decoder block

Additionally, to visually demonstrate the process of feature extraction and segmentation output by the improved LinkNet, Figure 6 selected one frame from multiple feature maps

extracted by each network layer for display. The sizes of the feature maps were standardized to facilitate detailed presentation. It can be observed that as the downsampling progresses, the resolution of the feature map gradually decreases, resulting in clearer semantic features but relatively vague positional information. During the upsampling process, the decoder network utilizes the obtained high-level semantic features to predict the category of each pixel. As the resolution of the feature map is progressively restored, the image details become richer, ultimately generating the segmentation result based on the number of categories.

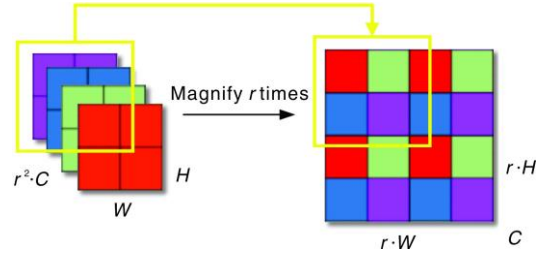


Fig. 5 Sub-pixel convolution

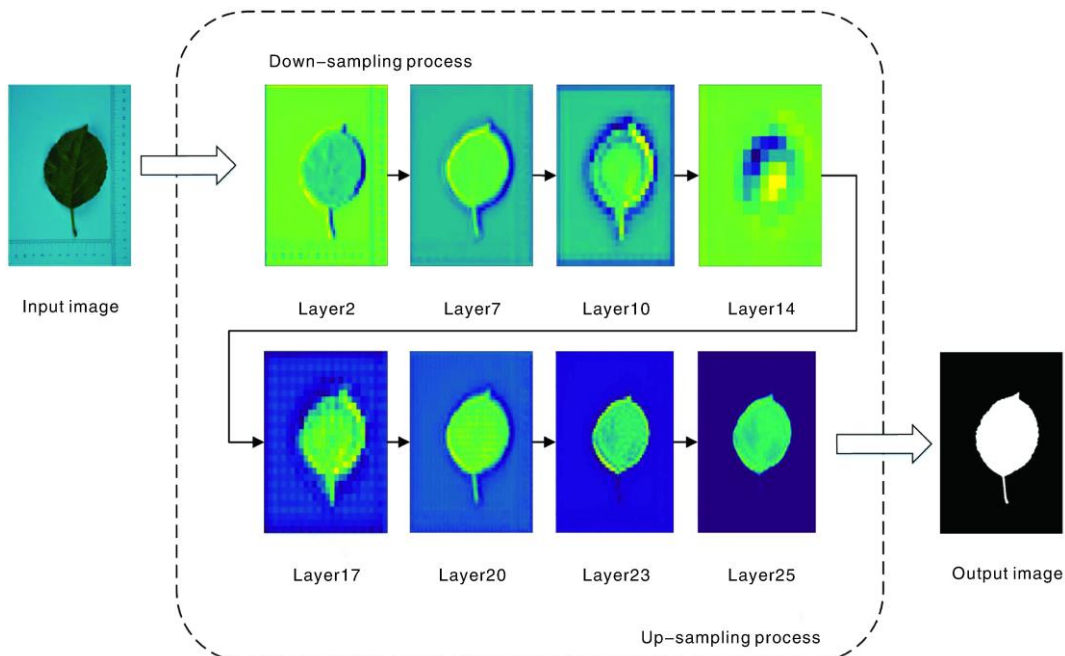


Fig. 6 Extracted feature maps of some layers of improved LinkNet

2.3. Experimental environment

The experimental environment is based on the PyTorch deep learning framework. The hardware setup comprises an i9-10900K CPU, an NVIDIA GeForce RTX 3090 GPU, and 64GB of RAM. The initial learning rate is set to 0.001, utilizing the Root Mean Square Propagation (RMSprop) optimization algorithm with an initial momentum of 0.9. The total number of iterations is set to 21.

The dataset for apple leaf images used in the experiment comprises a total of 8,184 images. Before the experiment, the dataset was divided into training and testing sets in a 7:3 ratio. Subsequently, 10% of the training set was extracted to create a validation set. As a result, the final training set consisted of 5,157 images, while the validation and testing sets comprised 572 and 2,245 images, respectively. Each dataset contains 174 classes of apple leaf images proportionally. The training set utilized the Focal Loss function [11].

2.4. Evaluation Index

The objective of leaf image segmentation is to classify each pixel in the image, thereby distinguishing the main leaf subject from the background. To evaluate the segmentation performance of the model, this study employed two metrics: Mean Pixel Accuracy (mPA) and Mean Intersection over Union (mIoU). The mPA metric represents the average percentage of correctly predicted pixels for all categories compared to the total pixels in each category. On the other hand, mIoU represents the average ratio between the intersection and union of predicted and ground truth values for all categories.

In order to better assess the model performance and demonstrate its practical applicability, three additional metrics were selected in this study: Model Parameters (Paras), Floating Points of Operations (FLOPs), and Inference Time (the average time required for inferring a single image during testing). Parameters and FLOPs are used to measure the

model's size and computational complexity, respectively, while Inference Time provides an intuitive understanding of the network's inference speed. Additionally, Total Training Time, the duration required for the model to complete 21 training rounds, was utilized to evaluate the impact of pre-trained models on training time.

3. Results and Analysis

Firstly, based on the original LinkNet, the main backbone networks of the encoder utilized pre-trained models ResNet18 and ResNet34, as depicted in Table 1 with the experimental results. Comparing experiments 1 and 2 indicates that using ResNet18 as the encoder's main backbone network reduced

the total training time by 1.4 hours and decreased the inference time by 1.01 milliseconds. Additionally, the model's parameter count and floating-point operations slightly decreased. Contrasting experiments 1 and 3 shows that adopting ResNet34 as the encoder's main backbone network shortened the total training time by 1.34 hours and improved segmentation accuracy by 0.54 percentage points. However, due to its more complex network structure, it increased the model's parameter count and floating-point operations, consequently extending the network's inference time by 3.38 milliseconds. Hence, employing pre-trained models ResNet18 and ResNet34 significantly reduced the model's training time, with comparable segmentation accuracy across all experiments.

Table 1. Comparison of experimental results with the introduction of different pre-trained models

Test No.	Backbone	mIoU/%	mPA/%	Total training time/h	Inference time	Paras/M	FLOPs/G
1	—	96.15	98.01	4.82	13.01	11.54	43.69
2	ResNet18	96.18	98.74	3.42	12	11.53	43.64
3	ResNet34	96.69	98.93	3.48	16.39	21.64	78.6

Based on the experimental results in Table 2, various ablation experiments were conducted using the LinkNet with ResNet18 as the backbone network as the baseline, exploring different improved structures. The test results on the apple leaf image test set for different improved structures are presented in Table 3. Experiment 1 represents the usage of pre-trained model ResNet18 as the backbone network in LinkNet. Experiment 2 depicts a reduction in the number of encoder-decoder pairs from 4 to 3, based on Experiment 1. Experiments 3, 4, and 5 correspondingly represent different scale improvements in the channel reduction scheme based on Experiment 1, while Experiment 6 signifies the substitution of sub-pixel convolution for deconvolution in the final block of Experiment 1. Experiment 7 showcases the combined results of introducing all the improved structures into the enhanced LinkNet.

Comparing Experiments 1 and 2 reveals that reducing the encoder-decoder pairs from 4 to 3 decreased the inference time by 1.18 ms. This reduction is attributed to maintaining a smaller depth and width in the network by appropriately reducing the number of encoder-decoder pairs, significantly reducing the parameter count and hence the inference time. Contrasting Experiments 1, 3, 4, and 5 shows that improving the channel reduction scheme also accelerates the

network's inference speed. The m/16 channel reduction scheme demonstrates the best comprehensive performance. Compared to m/8, it offers fewer parameters and shorter inference time while maintaining comparable accuracy. Compared to m/32, it exhibits higher accuracy with minimal changes in parameter and floating-point operation counts.

Experiment 4, altering the channel reduction scheme to m/16, reduced the inference time by 0.36 ms with a marginal 0.02% decrease in accuracy compared to Experiment 1. Comparing Experiments 1 and 6 highlights that replacing the final block with sub-pixel convolution shortened the inference time by 2.36 ms, significantly speeding up the inference process. Sub-pixel convolution eliminates the need for learning parameters, substantially reducing both the floating-point operation count and parameters during upsampling.

Therefore, as seen in Table 2, these various improved structures effectively reduce the parameter count and floating-point operation count produced during model inference, shortening the network's inference time with minimal impact on segmentation accuracy. Among these improvements, reducing the number of encoder-decoder pairs and sub-pixel convolution demonstrate the most noticeable enhancement in inference speed.

Table 2. Ablation studies design and comparison of the results

Test No.	Different improved structures			mIoU/%	mPA/%	Inference time/ms	Paras/M	FLOPs/G
	Encoder-Decoder blocks	Channel reduction	Final block					
1	4	m/4	—	96.18	98.74	12	11.53	43.64
2	3	m/4	—	96.02	98.04	10.82	2.89	35.14
3	4	m/8	—	96.17	98.12	11.75	11.32	41.97
4	4	m/16	—	96.16	98.02	11.64	11.25	41.47
5	4	m/32	—	95.45	97.54	11.56	11.24	41.36
6	4	m/4	Sub-pixel conv	95.42	97.41	9.64	11.52	37.89
7	3	m/16	Sub-pixel conv	95.01	97.27	7.82	2.81	27.88

Comparing Experiment 1 in Table 2 and Experiment 7 in Table 3 reveals that using the improved LinkNet for leaf segmentation resulted in an mIoU of 95.01%. This figure is

only 1.14% lower than the original network. However, the model's parameter count decreased by 75.65%, and the floating-point operation count reduced by 36.19%. The

inference speed increased by 39.89%, with an average inference time of only 7.82 ms per image. This reduction of 5.19 ms compared to the original network's inference time meets practical application requirements.

4. Conclusion

This study introduced four improvements to the LinkNet model: incorporating pretrained models, removing a pair of encoding-decoding blocks, enhancing channel reduction, and utilizing sub-pixel convolutions. All these enhancements reduced training time, accelerated model inference, and improved segmentation efficiency. Although it incurred a certain loss in accuracy, the model size notably decreased. Additionally, from the segmentation results, despite a slightly lower segmentation accuracy, the leaf edges remained sharply defined, demonstrating that the improved LinkNet is suitable for practical applications.

The proposed method exhibited promising results on leaf images. Further enhancements could involve augmenting the training dataset with more natural leaf samples to enhance the model's generalization capabilities. This research approach offers a new direction for efficiently and accurately measuring leaf geometric parameters in open environments. Moreover, it lays the groundwork for embedding lightweight segmentation models into embedded devices, paving the way for low-cost, high-precision, and fast leaf area measurement equipment.

References

- [1] YU D Y, FENG T X, LI Y X, et al. Research and implementation of living leaf area measurement based on plant image[J]. *Intelligent Computer and Applications*, 2019,9(4) : 173 - 176.
- [2] LIN K, CHEN J, SI H, et al. Measurement of leaf area index using reference object and computer vision[J]. *Oxidation Communications*, 2016, 39(2A) : 1832 - 1844.
- [3] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4) : 640 - 651.
- [4] GAO M, VAN DER HEIJDEN G W A M, VOS J, et al. Estimation of leaf area for large scale phenotyping and modeling of rose genotypes[J]. *Scientia Horticulturae*, 2012, 138: 227 - 234.
- [5] PAN S J, YANG Q. A survey on transfer learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22 (10) : 1345 - 1359.
- [6] YU C Q, WANG J B, PENG C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation[C]//*Computer Vision -ECCV 2018*, 2018.
- [7] CHAURASIA A, CULURCIELLO E. LinkNet: exploiting encoder representations for efficient semantic segmentation[C]//*2017 IEEE Visual Communications and Image. Processing*. St. Petersburg, FL, USA. IEEE, 2017.
- [8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA. IEEE, 2016 : 770 - 778.
- [9] EASLON H M, BLOOM A J. Easy Leaf Area: automated digital image analysis for rapid and accurate measurement of leaf area[J]. *Applications in Plant Sciences*, 2014, 2 (7) : 1400033.
- [10] Zhang S, Zhang C. Modified U-Net for plant diseased leaf image segmentation[J]. *Computers and Electronics in Agriculture*, 2023.
- [11] A A T, A B B, B J G, et al. Deep Leaf: Mask R-CNN based leaf Detection and Segmentation from digitized herbarium specimen images[J]. *Pattern Recognition Letters*, 2021.