Computers and Electronics in Agriculture 193 (2022) 106712

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

A single view leaf reconstruction method based on the fusion of ResNet and differentiable render in plant growth digital twin system

Wei Li [a,b], Deli Zhu [a,b,*], Qing Wang [a,b]

[a] College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
[b] Chongqing Digital Agricultural Service Engineering Technology Research Center, Chongqing 401331, China

ARTICLE INFO

Keywords:
Deep learning
3d leaf reconstruction
Resnet
Differentiable render
Plant growth digital twin system

ABSTRACT

In modern agriculture, plant growth digital twin system helps breeders monitor plant growth, increase yield, and provide growth management advice. Research on the single view leaf 3D reconstruction in digital twin systems has achieved relative success. However, in traditional single-view reconstruction algorithms, the leaf reconstruction often contains the problems of low precision, achieving complexity, and slow speed, making it difficult for recovering three-dimensional information about leaves. Consequently, the reconstruction precision is significantly reduced, which further affects the accuracy of single-view leaf 3D reconstruction. In response to this problem, this study proposed a single-view leaf reconstruction approach in plant growth digital twin systems based on deep learning. The method in this paper mainly fuses the advantages of ResNet and differentiable rendering, and the model is used for further enhancing feature extraction capability and reconstruction precision. Finally, the experiment presented in this paper suggests that the method allows for the 3D reconstruction of plant leaves with different shapes using a single view. Moreover, the experiment results show that the F-Score, CD, EMD reached 76.192, 0.808, and 3.567. Compared with other models, the proposed model in this study has higher reconstruction accuracy, 3D evaluation indicators, and prediction results, providing important ideas and methods for recovering the leaves from a single view in a plant growth digital twin system.

1. Introduction

In the current field of smart agriculture, the leaf is the organ that produces nutrients and is the basis for crop growth and development, therefore, leaf reconstruction has great research significance in crop growth monitoring, however, the digital twin system for plant growth not only quantitatively describes the physiological mechanisms of leaves, growth simulation but also responds to the complex surroundings environment. Leaf reconstruction is the task of analyzing plant growth traits to describe plant growth in a digital twin system. Also, breeders use the 3D geometry of plant leaves to support decisions in plant fields. Such decisions include selecting the best cultivars to continue the breeding process and selecting the best cultivars for the following seasons (Zhao et al., 2010). Therefore, a rapid and accurate leaf reconstruction method is essential for the analysis of leaf phenotypes, parameter measurements, and virtual visualization in the digital twin system for plant growth (Vos et al., 2010; Guo et al., 2011).

Traditional leaf reconstruction methods in single view mainly rely on the point cloud, shapes, and shadows, etc. Mundermann et al. (2003) have implemented a single-image 3D reconstruction of plant leaves based on the extraction of leaf silhouette, calculation of the leaf vein skeleton, and rotational changes; Tan et al (2008) used an interactive method to segment the branches and crowns of trees, and then automatically generated 3D tree leaves and branches based on models in the model library; Wang (2013) used orchard leaves as the research object and constructed a 3D reconstruction algorithm for single images based on surface light and dark changes. Ren Feier et al (2021) used the method of shadow recovery to reconstruct the three-dimensional shape of leaves from statistical brightness patterns and morphological features to achieve the three-dimensional reconstruction of single-image leaves in the field.

Although the above-mentioned studies on disease recognition have achieved ideal results, there are large uncertainties in the reconstruction performance and poor accuracy and stability of the algorithm. In recent years, there has been an increase in studies employing single view reconstruction in deep learning fields. According to Zhu et al. (2020), a large number of network models have been developed for different 3D model representations, each with its advantages and disadvantages.

* Corresponding author.
E-mail address: 463453339@qq.com (D. Zhu).

https://doi.org/10.1016/j.compag.2022.106712
Received 31 October 2021; Received in revised form 24 November 2021; Accepted 11 January 2022
Available online 22 January 2022

Choy et al. (2016) proposed a 3D-R2N2 network based on voxel-based deep learning 3D reconstruction, which includes an encoder, a 3D LSTM neural network, and a decoder, and integrates a single view and multiple views in a single framework to achieve better end-to-end 3D reconstruction. However, the resolution of the model output is only 323, and the accuracy of the model reconstruction is not high. Given the deficiency of voxels, Fan et al. (2017) proposed a deep learning reconstruction method based on the 3D point clouds, which designed a conditional sampler to estimate the spatial information of the real point clouds from a given image, and Tatarchenko et al. (2017) proposed a loss function of EMD and CD synthesis determination to improve the reconstruction accuracy of the model, but the 3D point cloud is visually difficult to achieve real-world effects. Wang et al. (2018) proposed a mesh reconstruction method, which extracts image feature results by a CNN network, and uses a GCN convolutional network to store 3D spatial information, and deforms 3D meshes based on cascaded deformation modules to solve the problem of 3D mesh inverse pooling. However, these reconstructed networks are based on open datasets, and the reconstructed shapes are all relatively regular object models. Chen et al. (2019) and Liu et al. (2019) proposed a 3D object reconstruction using an interpolation-based differentiable render from a single 2D image.

Although the research on leaves reconstruction in traditional methods and deep learning has made some progress, fewer works exploit deep learning to complete 3D plane leaf geometry, which is important to evaluate plant growth in a digital twin system. Based on the existing research, this study tackles reconstruction in a single view, through an end-to-end leaves reconstruction by fusing ResNet and differentiable render in a single view. The main contributions of this study are as follows:

(1) In this paper, a leaf reconstruction model with a single view was proposed based on deep learning for RGB images with multi-views in rendering. This model achieved precise reconstruction for leaves and provided a basis for building a digital leaf model.
(2) By improving the network structure of ResNet18 as a feature extraction structure for the reconstruction results, this method can provide better feature extraction capabilities for precise reconstruction in single-view.
(3) A differentiable rendering method for optimizing the silhouette shape of leaf mesh. By calculating 2D image loss between the input image and rendered prediction, the model goal is to deform the 3D shape of the prediction mesh.

2. Materials and methods

2.1. Data description

This paper mainly focused on single-view leaf reconstruction in deep learning using self-collected datasets. The neural network input consists of point clouds with normal vectors and RGB images, both generated from the set of leaf models in this paper. Taking into account the accuracy of reconstruction from a single view, leaf models were collected to obtain high precision and low noise in different shapes and sizes. The categories of leaf models collected included ellipsoid leaves, heart-shaped leaves, truncate leaves, oval leaves, round leaves, etc. Examples of the selected images are shown in Fig. 1.

2.2. Data processing

A total of 9 different shape leaf models in this study(including the above leaves shape), and this study has to convert the mesh into multi-view rendering RGB images and point clouds as input data of the neural network. First, the RGB images were obtained by setting the multi-view camera position to render the normalized mesh and resized to 224×224 pixels. Then, the leaf point clouds were sampled by mesh and calculated the vertex normal vectors (for the calculation of normal loss in the network). In this paper, the 20,736 images were divided into the training set and test set respectively according to the radio of 4:1. The RGB images and point clouds labels are shown in Fig. 2.

2.2.1. Data normalization

The source of the training data in this paper is the leaf model, which is obtained by the 3D modeling method. In order to ensure the quality of the rendered view, the initial model is normalized as shown in formula (1), which is the coordinates of the model in 3D space, and the minimum and maximum values of the coordinates of each dimension respectively, and the model dimensions are scaled in a cube enclosing the box of the specified size, while the center of the model is set to the origin and the model coordinates are controlled to be between −1 and 1.

$$p_i^{'} = \frac{p_i - mean(p_i)}{max(p_i) - min(p_i)}, i \in (1, 2, ..., n) \tag{1}$$

where p_i is all vertex position of the mesh in datasets and $p_i^{'}$ is the normalized p_i. In this paper, the normalization of each vertex in the mesh reduces the error caused by different units and improves the efficiency of model rendering and network training.

2.2.2. Point cloud processing

Point clouds are an important part of the dataset. However, the inconsistent number of vertices or faces of the acquired 3D models, which is not conducive to network training convergence. Therfore, point clouds were processed before training. Firstly, this study uniformly sampled a fixed number of 1000 point clouds from the original mesh with the Meshlab. Then, due to the need to calculate the normal loss function, the normal vector of point clouds was calculated based on the sampled vertices. The image after point clouds processing is shown in Fig. 3.

2.2.3. Rendering

In this paper, referring to the literature of Choy et al.(2016) for the rendering method, this study achieves the task of rendering the leaf mesh model by setting camera positions in the scene. At the same time, different camera angles and light intensity are combined to render the leaf model of the scene, and the rendered view of different angles can better simulate the image in the real world. Fig. 4 shows the rendering results of different camera positions. This method is able to reconstruct the 3D model from a single image of the leaf because the reconstruction network can learn the mapping relationship between the 2D image features of the leaf in different views and the 3D mesh vertices during the training process, and after fitting, it can predict the invisible 3D spatial parts from a single image.

2.2.4. Perspective projection

Wang et al. (2018) mentioned in the literature how the key to ensuring that the correct link can be made between the two different

Fig. 1. Example of leaf models.

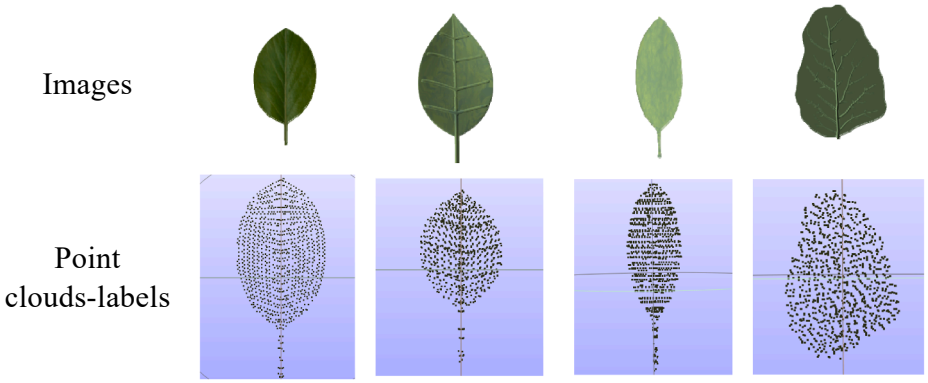

Fig. 2. Image and point clouds.

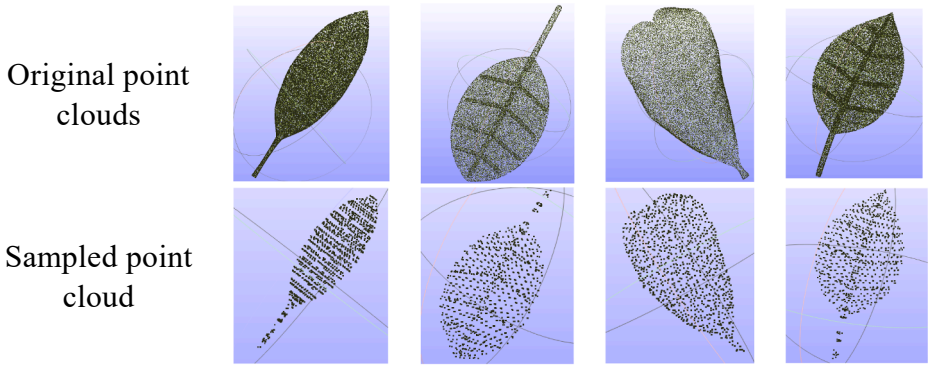

Fig. 3. Point clouds processes.

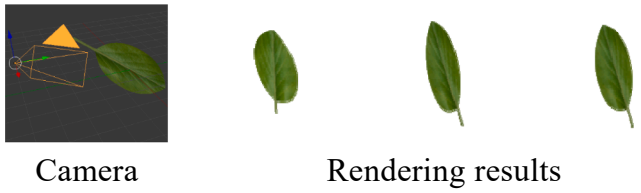

Fig. 4. Rendering results from different camera perspectives.

forms of data and to better utilize the 2D image information to help reconstruct the 3D mesh model is to map the coordinate information corresponding to the 3D points of the model obtained back into the rendered view based on the camera parameter information previously used for rendering. As shown in Fig. 5, it is ensured that the generated mesh vertices correspond in projection to the 2D feature points in the rendered view. As shown in Fig. 5, ensure that the resulting mesh vertices correspond to the 2D feature points in the rendered view in projection. Projection matching is to avoid the deviation of two-dimensional image information from 3D coordinate information, resulting in mismatches between the reconstructed model and the input image, position deviations, etc.

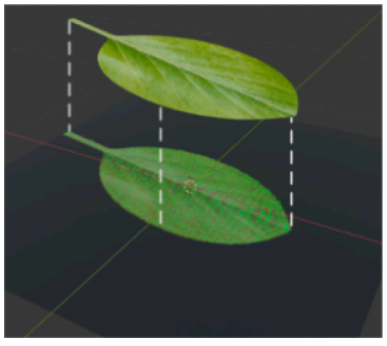

Fig. 5. Perspective projection.

2.2.5. Data enhancement

Deep learning requires sufficient data to complete the training process. Increasing the size of the dataset appropriately is conducive to the improvement of reconstruction accuracy. Therefore, in this research, a total of 9 mesh of leaves were enhanced by two methods. The first way is by random angle rotation, where the mesh is rotated at a randomly selected angle between 0 and 180 to obtain the enhanced new image and point cloud. The second way is by random position panning, where a value between −1 and 1 is randomly chosen to adjust the position of the mesh, giving an enhanced image and point cloud.

2.3. An overview of single view leaf reconstruction model

Fig. 6 shows an overview of the proposed model, single-view leaf reconstruction model, which consists of three stages: (1) the data preprocessing stage (Section 2.2), (2) the feature extraction stage, and (3) the learning of the fused image features and 3D features and predicting the leaf mesh stage. (4) the refined model with a differentiable render stage.

In stage one, the model has three main tasks: (1) processing the original image data, (2) preprocessing the original poly data of mesh. (3) rendering images. For the input image data, the images are generated based on specific camera positions (Section 2.2.3) rendering. For the original point clouds, the model downsamples and normalizes the point cloud data (Section 2.2.2). Finally, the preprocessed image and point clouds data are used as inputs for stage two.

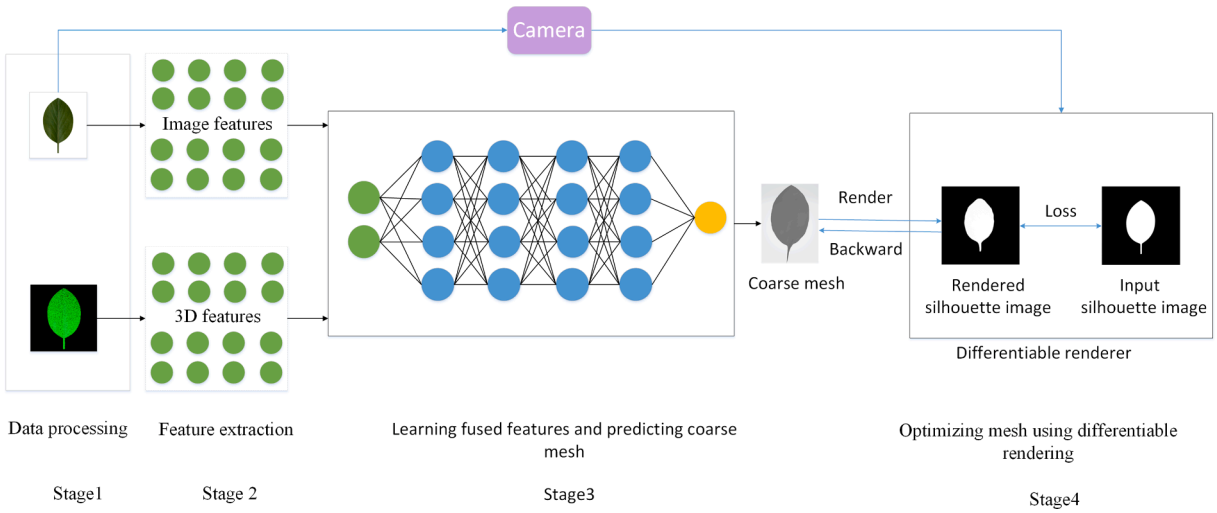

Fig. 6. An overview of single view leaf reconstruction model.

In stage two, the improved deep learning networks of ResNet are used to capture image features from all of the images preprocessed in datasets. Vertex positions and normal vectors are extracted from the preprocessed point cloud dataset based on Meshlab. Stage three of the network learns and converges image features and 3D features. The stage focuses on extracting features from images and 3D data for the next stage

In stage three, the model (Pixel2Mesh) first combines these two kinds of deep features by fusing the output of ResNet and 3D features of point clouds and learns the relationship between the two kinds of deep features for generating the mesh by GCN. Finally, the model predicts leaf-shaped mesh from RGB images.

In stage four, this paper uses the differentiable rendering method to deform the coarse leaf mesh by comparing the silhouette between the rendered image and the target image. With the differentiable renderer, the goal is to optimize the coarse mesh to fit the current input image in a single view.

2.4. Learning image features using ResNet

In deep learning, ResNet (He et al., 2019) is a class of convolution neural network(CNN) that applies the shortcut connection of residual module, which has the advantage of avoiding gradient loss or explosion and precision degradation caused by simple stacking convolutional neural networks and enhance the transmission of image features without increasing parameters and computational complexity.

As shown in Fig. 7, assuming that the optimal solution map is represented by H(X) = x, the general convolution network is fitted directly by the above equation, while the residual network expects the residual mapping to be fitted, i.e. F(X) = H(X)-X, where the optimal solution of the network is H(X) = F(X) + X. From this equation, it can be found that the residual network will automatically extract features only when F (X) = 0, i.e. The constant mapping H(X) = X is completed, and it is easier to learn F(X) than to learn H(X).

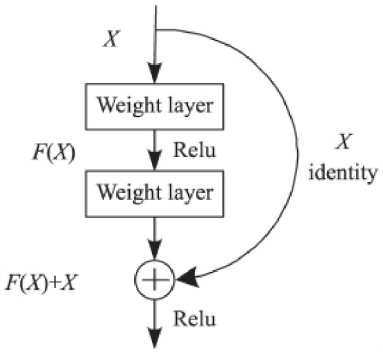

Fig. 7. The basic structure of the residual network.

Wang et al. (2018) used a neural network similar to VGG-16 as the feature extraction module in the Pixel2Mesh method, however, VGG has limited ability to extract feature information compared to ResNet, and it is prone to imperfect and in detail feature extraction for single-view images of complex leaves in the dataset of this paper. Secondly, the residual network can effectively solve the degradation problem caused by increasing the depth and improve the performance of the network by simply increasing the depth of the network, which is a more easily optimized network. In addition, since deep learning network training requires a large amount of data, overly tiny datasets is prone to overfitting. In order to prevent overfitting, the simpler the model, the better the fitting effect for smaller data sets. In this paper, the network structure was selected by comparing the performance of common 18-layer, 34-layer, and 50-layer networks on 3D reconstruction, and finally, ResNet18 was selected, whose structure is shown in Fig. 8, and the network was improved and optimized according to this network.

ResNet18 has more applications and great results in the field of image recognition and segmentation, but most of the ResNet18 initially extracts input image features with a 7 × 7 convolutional kernel in 64 dimensions, although the large convolutional kernel has a larger perceptual field and can better extract the features of images with the larger size. For single-view 3D reconstruction, the main construction is the mapping relationship between the image and 3D vertices, the more detailed features of the image, the more accurate the three-dimensional mapping relationship will be built, in this training dataset of paper rendering view with many different angles and small outlines, the use of the above method will lead to the reconstruction network in the initial study of the lack of shallow two-dimensional feature information, loss of some image features, thereby affecting the accuracy of 3D reconstruction (Zhuang, 2020).

To solve the above network problems, The first two layers of ResNet18 are added with 3 × 3 16-dimensional convolutional kernels and 32-dimensional convolutional kernels respectively, which are used to extract 2D shallow image feature information; Secondly, The purpose of replacing the 64-dimensional 7 × 7 convolutional kernels in the original ResNet18 with 3 × 3 small convolutional kernels is to obtain more detailed features by using smaller-scale convolutional kernels. Fig. 9 shows the structure of the improved s feature extraction network.

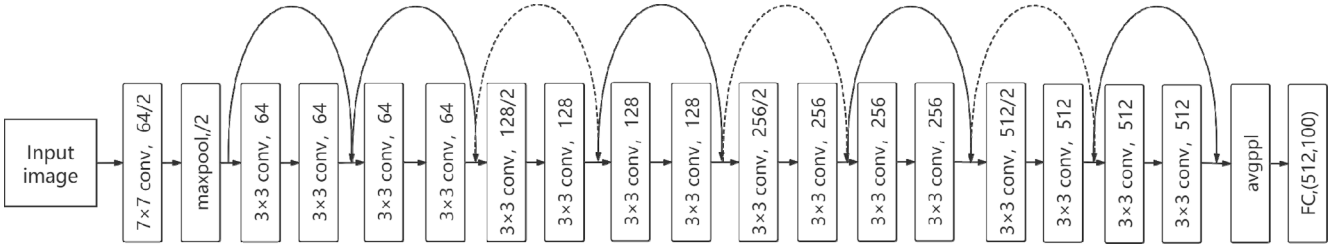

Fig. 8. Structure diagram of resnet18.

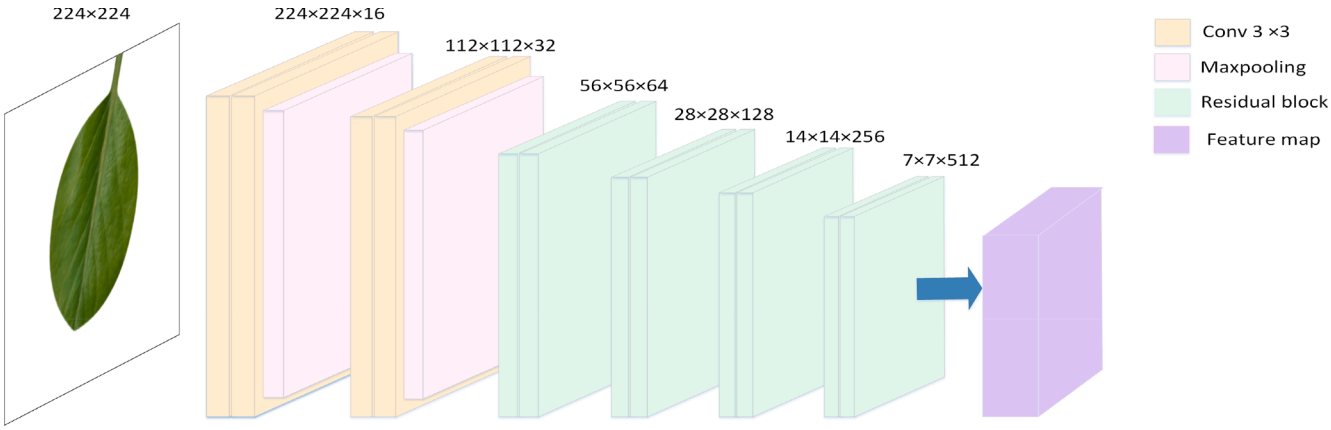

Fig. 9. Structure diagram for feature extraction network.

2.5. *Learning fused features using Pixel2Mesh*

Pixel2Mesh (Wang et al., 2018), an end-to-end 3D reconstruction neural network from a single-view image, optimizes some of the shortcomings in traditional single-view reconstruction algorithms. voxel reconstruction network (Choy et al. 2017), point clouds reconstruction network (Fan et al., 2016) and solves generating 3D Mesh Models from Single RGB Images. Based above reason, a network incorporating ResNet and Pixel2Mesh was used to learn deep features from 2D features and 3D features and predict leaf mesh.

The 3D reconstruction network shown in Fig. 10 is composed of two parts: the upper part is the feature extraction module, which extracts the leaf image features at different levels layer by layer, and the lower part is designed with three cascading deformation modules, composed of GCNs, which are used to store the mesh information after each deformation. The initial mesh (ellipsoid) is continuously deformed by the 2D feature information from the perceptual feature pooling layer to approximate the real leaf 3D shape, and each deformation process fuses the leaf image feature points with the vertex features in the GCN as a new round of input to the GCN, thus updating its 3D vertices and features, At the same time, the number of nodes and edges is increased by the graph unpooling layer to increase the level of detail of the model. Finally, the initial model is then continuously transformed to the target model through iterative updates.

To optimize the above-mentioned improved reconstruction network has a better effect, this paper adjusts the network parameters by minimizing the difference between the real network and the reconstruction grid. For the loss function of the network, this paper still uses the loss

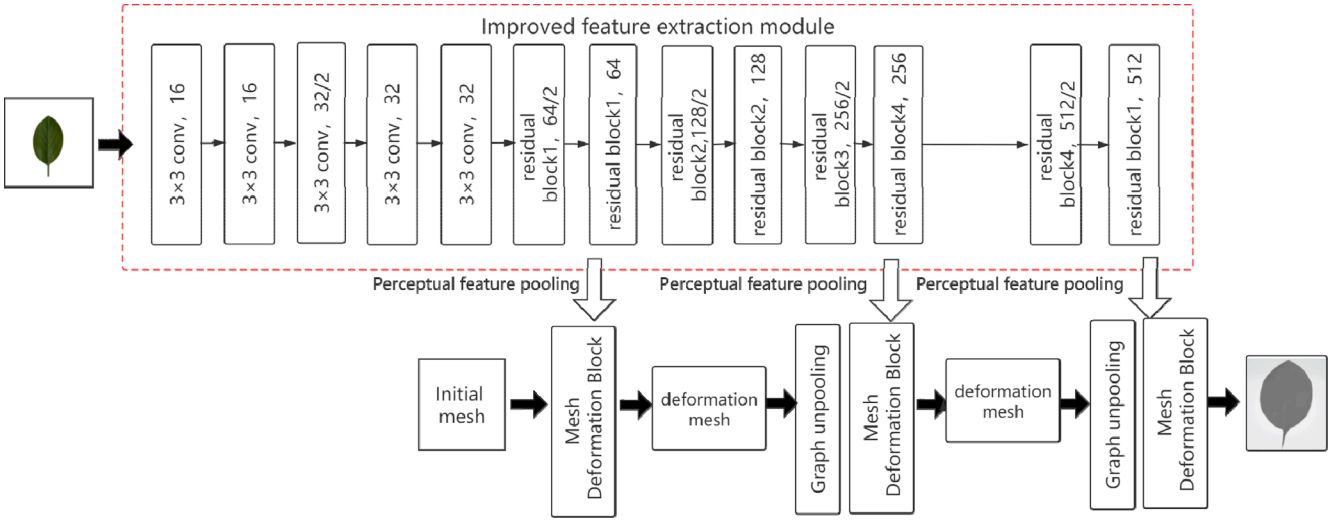

Fig. 10. Single view 3D reconstruction network structure diagram.

function in the paper (Wang et al., 2018) and defines the loss function of the neural network from both vertex and method vector, including the following four functions:

The Chamfer loss function, also known as chamfer distance, is designed for each vertex in one collection and looks for the nearest point in another collection, adding its square distance to verify that the vertices of the 3D mesh are similar to each other in the actual 3D mesh and that the purpose is to limit the specific position of the grid vertices.

$$l_c = \sum_p \min_q ||p-q||_2^2 + \sum_q \min_q ||p-q||_2^2 \quad (2)$$

Among them, $\boldsymbol{p}$, $\boldsymbol{q}$ represents the vertex matrix of the 3D prediction network and the real network respectively.

Normal loss function, reconstructing the network not only considers the vertex position of the 3D mesh but also considers the loss of features of the mesh faces to ensure the consistency of the mesh surface normal vectors to increase the smoothness of the surface.

$$l_n = \sum_p \sum_{q=argmin_p(||p-q||_2^2)} || < p-k, n_q > ||_2^2 \quad (3)$$

where the $\boldsymbol{q}$ is the closest vertex to p calculated by the Chamfer function; and $\boldsymbol{k}$ is the point adjacent to p, n_q represents the normal vector of the vertex $\boldsymbol{q}$ to the vertices in the real grid.

Laplacian regularization, a function that serves to maintain the relative position of adjacent vertices during deformation. The $\delta_p^{'}$ and δ_p in formula (4) represent the adjacent points and sub-adjacent of vertex p, respectively.

$$l_{lap} = \sum_p ||\delta_p^{'} - \delta_p||_2^2 \quad (4)$$

Edge length regularization, which is designed to prevent local optimization of individual anomalies, resulting in uneven mesh surfaces, as shown in formula (5).

$$l_{loc} = \sum_p \sum_k ||p-k||_2^2 \quad (5)$$

2.6. Refining coarse mesh with differentiable render

Given a coarse 3D mesh, such as predicted mesh in Pixel2Mesh, a differentiable rendering operator can be defined to compare the silhouette of current meshes to the silhouette of input images. With differentiable rendering, this study wants to generate images from a parametrized mesh that not only provides a rendered view of the mesh but also allows for differentiable with respect to the mesh vertices. In this way, this study can determine how the mesh should be deformed to match the desired input image in a single view. In this paper, the model uses the loss function of 3D mesh render like Liu et al. (2019).

Using differentiable rendering, such loss is minimized by defining a loss function in formula (6) and gradient descent, and machine learning methods are used to accelerate model optimization. This paper considers the predicted silhouette image generated by the render as $\hat{I}_s$ and its input silhouette image as I_s, and use an Intersection-Over-Union (IOU) loss as L_s for the silhouette prediction:

$$L_s = 1 - \frac{||I_s \otimes \widehat{I}_s||}{||I_s + \hat{I}_s - I_s \otimes \widehat{I}_s||} \quad (6)$$

where $\otimes$ denotes element-wise product. Once the differentiable renderer is defined, the model can use it to deform the coarse mesh by minimizing the norm between its rendered view and the target image, such that rendered image will be as similar as possible to the target image.

3. Experimental results and analysis

3.1. Model training

The hardware configuration used in this research for training and testing was shown in Table 1 follows. After trial and error, the hyper-parameters were determined as follows: the initial learning rate 3e-5, the number of epochs 50, batch size 1, and optimizer Adam. The leaves of all categories in the dataset were be shuffled out of order for training to prevent the network from overfitting and the training time was about 3 days.

3.2. Evaluation indicators

In this study, the standard 3D reconstruction indicators were used for evaluating the model proposed. An F-score Knapitsch et al., 2017 was calculated as a harmonic mean of precision, examining the percentage of points in prediction or ground truth that could find the nearest neighbor from the other within a certain threshold τ. Following Fan et al. (2017), Chamfer Distance (CD) and Earth Mover's Distance (EMD) were used for estimating model accuracy. For F-Score, larger is better. For CD and EMD, smaller is better.

3.3. Experimental results and analysis

3.3.1. Impact of the number of layers on the prediction performance

In deep learning, the number of layers in the network is an important consideration because it directly affects prediction performance. This section mainly aimed to compare the performance of the common 18, 34, 50 layers residual network used as feature extraction networks. After all of the testing data were scanned, the impact of the number of layers in the proposed model on the loss and reconstruction performance were evaluated. Fig. 11 demonstrated the loss values for the different number of layers (18 to 50) on the test set, and it could be seen from the results that the loss values on the test set decrease rapidly as the number of layers decreases.

In this paper, in contrast to determining the effect of different layers on the reconstruction from the perspective of loss, Fig. 12 showed that several random images from the test dataset were selected and input into the reconstruction network composed of different residual structures to check the effect of leaf model reconstruction, so as to determine which structure was more suitable for our single-view reconstruction network.

3.3.2. Comparison of 3D reconstruction indicators at different models

In this section, this paper use different 3D reconstruction metrics, such as F-Score, CD, EMD, to evaluate the effect of different reconstruction networks on the reconstruction results, using different models such as Pixel2Mesh (baseline). The comparison of results are shown below:

From Table 2, it was shown the F-scores for models with different thresholds, where $\tau = 10^{-4}$. Additionally, The CD and EMD for all leaf categories also were shown in Table 3.

On the other hand, the commonly used evaluation metrics for shape generation may not completely describe the leaf mesh quality. Such indicators tended to capture point-wise distance rather than surface properties, such as smoothness, continuity, surface detail. Thus, this paper also showed the results of the mesh shape in the next chapter,

Table 1
Experimental environment.

Experimental environment	Environment configuration
CPU	Intel Xeon Silver 411
Memory	64 GB
GPU	NVIDIA TITAN V
Deep learning framework	Tensorflow1.3.0

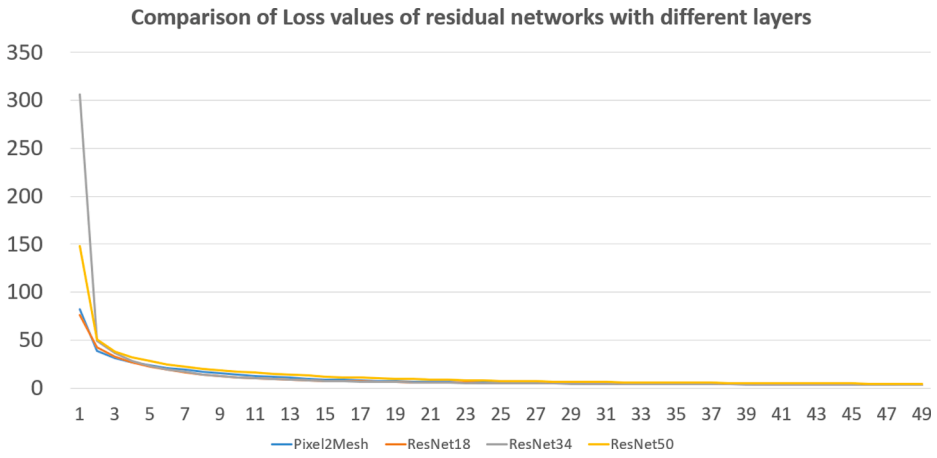

Fig. 11. Comparison of loss values of different number of layers of ResNet.

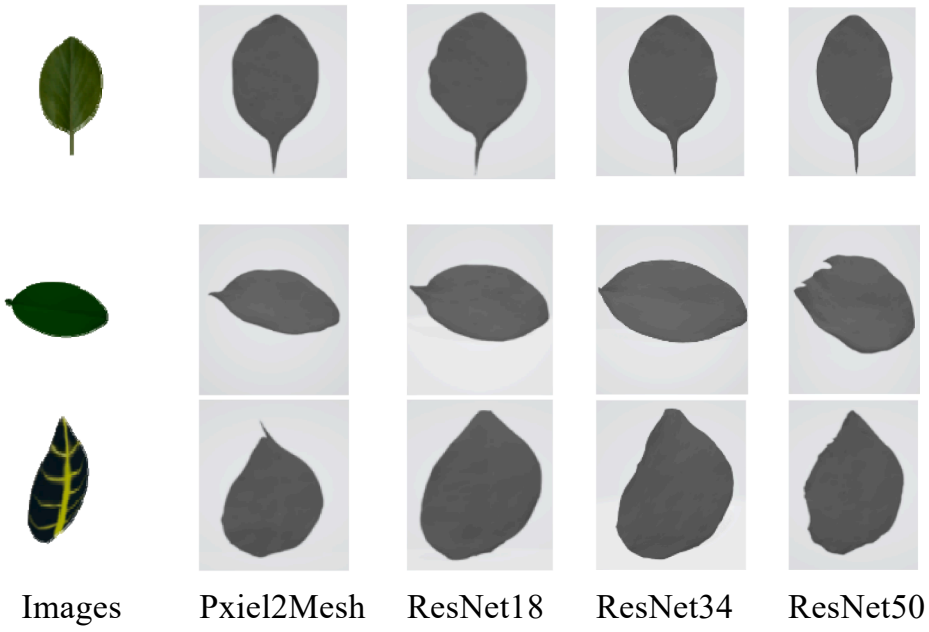

Fig. 12. Comparison of prediction results of different number of layers of ResNet.

Table 2
Comparison of F-Score indicators of different models.

Category	F-Score			
	τ		2τ	
	Pixel2Mesh	the model proposed	Pixel2Mesh	the model proposed
Ellipsoid	**73.934**	72.010	**83.342**	80.999
Truncate	93.164	**94.181**	98.332	**98.650**
Linear	85.020	**86.085**	90.905	**91.834**
Heart-shaped	41.278	**41.987**	54.341	**55.215**
Round	93.702	**94.281**	98.507	**98.810**
Long oval	48.966	**49.656**	58.392	**59.062**
Phialine	91.973	**93.526**	97.802	**98.416**
Oval	60.364	**61.097**	68.912	**69.470**
Tail Tip	**93.226**	92.903	**97.924**	97.911
Mean	75.736	**76.192**	83.162	**83.374**

Table 3
Comparison of CD indicators of different models.

Category	CD		EMD	
	Pixel2Mesh	the model proposed	Pixel2Mesh	the model proposed
Ellipsoid	**0.439**	0.529	**3.538**	3.715
Truncate	0.071	**0.065**	**1.149**	1.165
Linear	0.699	**0.503**	3.554	**3.458**
Heart-shaped	2.028	**1.806**	5.750	**5.615**
Round	0.076	**0.072**	2.455	**2.412**
Long oval	3.248	**2.484**	5.191	**4.797**
Phialine	0.070	**0.063**	3.152	**3.108**
Oval	**1.581**	1.683	**4.492**	4.513
Tail Tip	0.072	**0.071**	**3.060**	3.323
Mean	0.920	**0.808**	3.593	**3.567**

aiming to reflect the quality of the reconstructed mesh.

3.3.3. *Comparison of 3D reconstruction results at different models*

In this section of the experiments, the results of the reconstructed

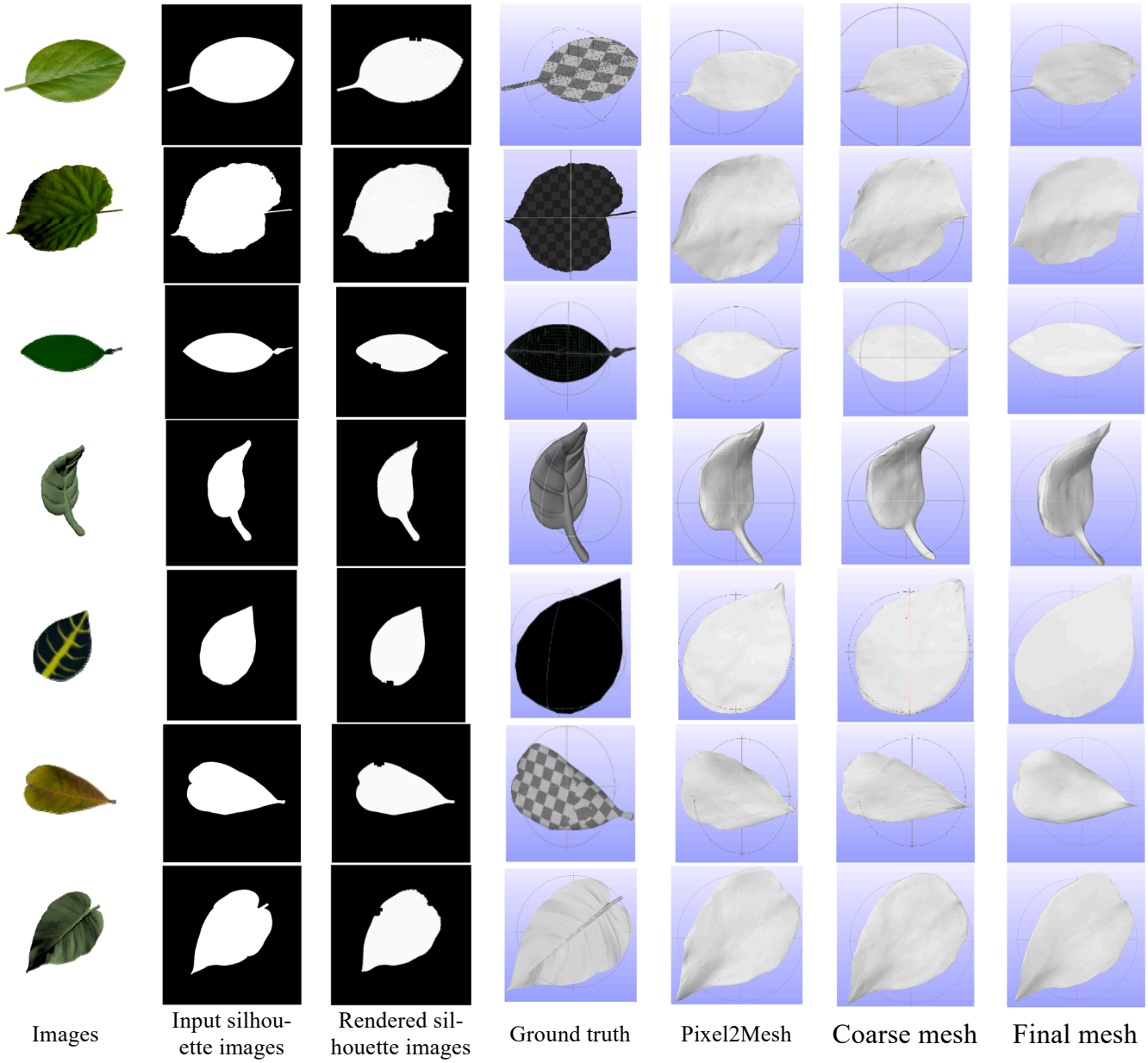

Fig. 13. Comparison of prediction meshes from partially shaped leaves of test datasets.

mesh of the methods on test data and realistic scenes were shown. Fig. 13 showed selected reconstruction results of plant leaf images from the test dataset. For each example, the paper showed the deformed mesh imposed over ground truth together with the target image, where the coarse mesh was the reconstruction result of the third stage and the final mesh was the final result after optimization of the model using differentiable rendering in this paper.

In this paper, to further illustrate that the method proposed could also be applied in real-world scenarios, a real-world plant leaf segmentation image as input and directly ran on the images, then compared to the baseline network as shown in Fig. 14. As can be seen, the approach generalized well to the real-world plant leaf images and provided a well-established virtual reconstruction method for the plant growth digital twin system.

4. Discussions

The contributions of the work in this study set out including the construction of datasets through traditional modeling methods, the use of improved residual convolutional networks to enhance the feature extraction capabilities of baseline networks, and the enhancement of the quality of mesh shapes based on differentiable renderers. In the experiment of results analysis(Section 4), it can be seen that the model of this paper has a significant reconstruction effect in different layers of residual network structure; compared with the benchmark network of Pixel2Mesh, the model proposed has certain improvements in 3D reconstruction index and reconstruction effect of test set; Correspondingly, in the evaluation of the reconstruction effect of a single-view leaf in the real scene, it can be seen that the reconstruction quality of the mesh, including the mesh shape, smoothness, etc., have more obvious improvement.

In the field of deep learning 3D reconstruction, compared to voxel (Choy et al., 2016) and point cloud (Fan et al., 2017) models, 3D models based on mesh representation have better realism and can map real-world leaves well (Wang et al. (2018)), while with the development of differentiable renderers, it makes single-view reconstruction better

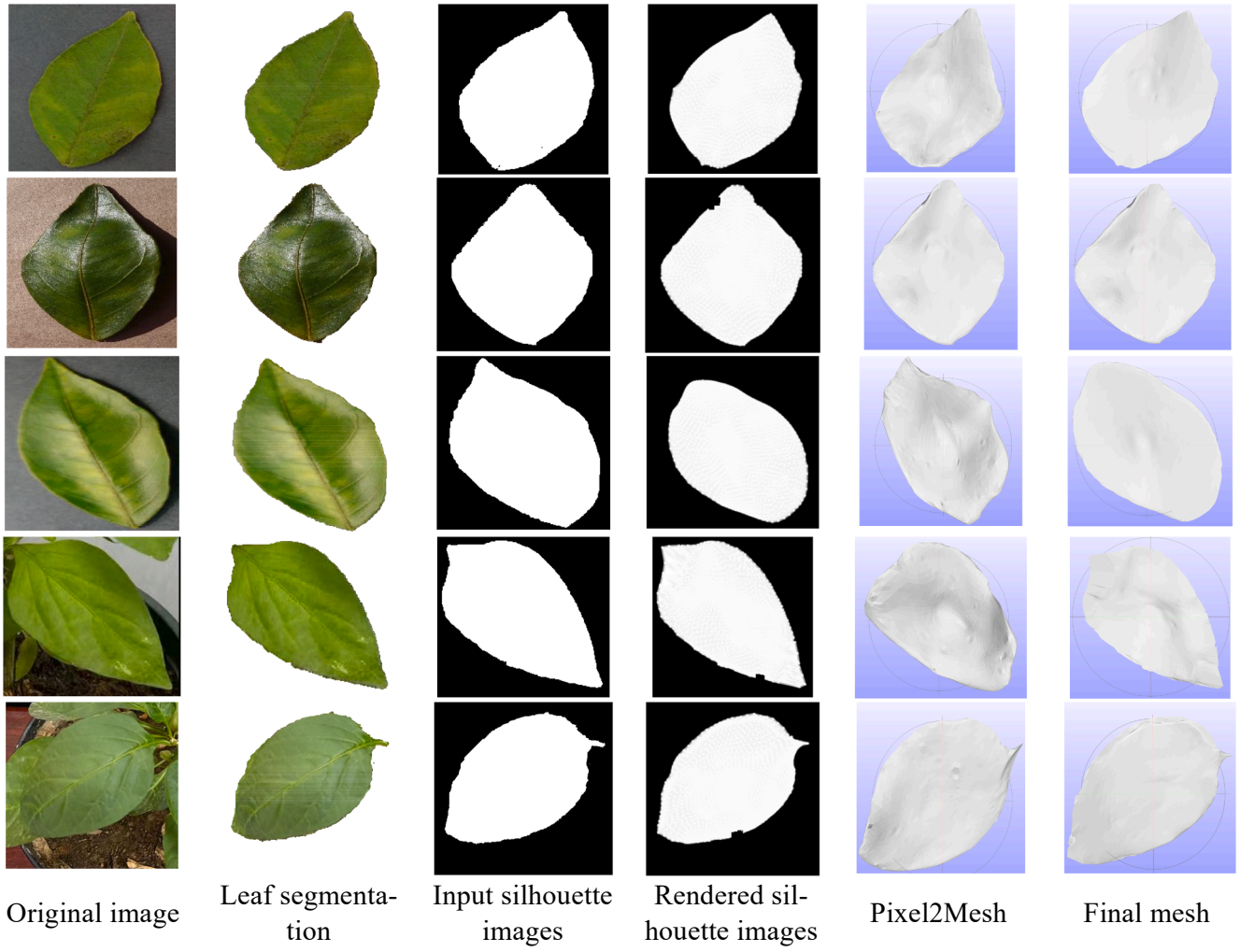

Fig. 14. Comparison prediction meshes from real-world plant leave images.

applied (Liu et al., 2019). However, the most of current single-view 3D reconstruction techniques based on deep learning are applied to object reconstruction, with less research on the digitization and visualization of leaves. The leaf is one of the important organs for measuring the growth status of plants. The advantage of this study is that by training the reconstruction network, a model database of leaves or other organs of plants can be constructed, and then using the target image, the source leaf mesh can be optimized and deformed using differentiable rendering techniques. Although the traditional 3D laser scanner can better restore the 3D structure of plants due to its high modeling accuracy, it is time-consuming due to the complexity of mechanical movements and operations; image reconstruction techniques can also restore better and effectively solve the time-consuming problem, but for the reconstruction of some crops with complex structures, a large amount of image data is required; (Gibbs et al., 2017, Gogoll et al., 2020, Zermas et al., 2017). Thus, based on the methods in this paper, morphological structures and organ models of plants can be better reconstructed using a single view, thus providing an effective reconstruction technique for plant growth digital twin systems and facilitating the effective monitoring of crop growth conditions in the system.

The model has achieved great results for predicting the mesh of single-view leaves based on an end-to-end method, however, the proposed method still has serval limitations to be solved:

(1) It is well known that deep learning training requires a large amount of data support to have good prediction results. In this paper, the data used for the experiments were collected by a relatively high precision method. However, the amount of data collected is slightly lacking compared to ShapeNet Chang et al., 1512 due to the lack of support from the equipment, and the shape of the leaves is more common. Therefore, the results of the network training also have some limitations, due to the insufficient dataset, the network has a weak generalization ability and can only identify some leaves with similar shapes.
(2) As the network structure and data inputs only contain an RGB image of a single leaf and a separate point cloud, the model is only capable of predicting a mesh of plane leaf from a single viewpoint and cannot reconstruct a model of the leaf from an image with stacked or multiple leaves.

In the future, the study in this paper will attempt to acquire more 3D models of different morphologies of leaves using high precision equipment and modeling methods. The images of the occluded parts and their reconstruction will then be extracted by machine learning or deep learning methods. In addition, this method will be applied to 3D vision tasks such as reconstruction and recognition detection of organs of different crops.

5. Conclusion

In this paper, a single view 3D leaf reconstruction method based on the fusion of ResNet and differentiable render was proposed for plant

growth digital twin system. The model uses a modified residual network to enhance feature extraction, then uses Pixl2Mesh to combine features and 3D positions of the image, and finally, refines the shape of the coarse mesh by differentiable rendering. The proposed model achieves 76.192 in F-score, 0.808 in CD, and 3.567 in EMD.

Based on the results of this paper, plants organ segmentation and extraction of, and refining the network structure of the model will be future research work aiming to improve network predictability. A limitation of this study is the small size of the plant dataset, therefore, future work will increase the number of samples to further improve the model precision. Overall, researchers can apply the model reconstruction techniques in this paper to visualization, leaf phenotype monitoring and analysis. The contribution of this paper can not only improve the efficiency of crop modeling and reduce the time and effort spent on modeling in terms of 3D reconstruction techniques in smart agriculture, but it can also be applied to crop target detection, segmentation of crop organs, and 3D visualization interactions based on 3D data, thus enabling effective construction of mapping between the 2D and 3D worlds.

CRediT authorship contribution statement

Wei Li: Writing – original draft, Software, Validation. **Deli Zhu:** Writing – review & editing, Supervision. **Qing Wang:** Data curation.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by the Science and Technology Research Project of Chongqing Municipal Education Commission (No. KJQN201800536); Research Project on Machine Vision Perception and Intelligent Algorithm for Intelligent Agriculture of Chongqing University Innovation Research Group (No. CXQT20015).

References

Chang A X, Funkhouser T, Guibas L, et al. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015.

Chen, W., Ling, H., Gao, J., et al., 2019. Learning to predict 3d objects with an interpolation-based differentiable renderer. Adv. Neural Information Processing Systems 32, 9609–9619.

Choy, C.B., Xu, D., Gwak, J.Y., et al., 2016. 3d–r2n2: A unified approach for single and multi-view 3d object reconstruction[C] //European conference on computer vision. Springer, Cham, pp. 628–644.

Fan H, Su H, Guibas L J. A point set generation network for 3d object reconstruction from a single image[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 605-613.

Gibbs, J.A., Poundl, M., French, A.P., Wells, D.M., Murchie, E., Pridmore, T., 2017. Approaches to three-dimensional reconstruction of plant shoot topology and geometry. Funct. Plant Biol. 44 (1), 62–75.

D. Gogoll, P. Lottes, J. Weyler, N. Petrinic, and C. Stachniss. Unsupervised Domain Adaptation for Transferring Plant Classifification Systems to New Field Environments, Crops, and Robots. In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS),2020. D. Gogoll, P. Lottes, J. Weyler, N. Petrinic, and C. Stachniss. Unsupervised Domain Adaptation for Transferring Plant Classifification Systems to New Field Environments, Crops, and Robots. In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS),2020.

Guo, Y., Fourcaud, T., Jaeger, M., et al., 2011. Plant growth and architectural modeling and its applications. Ann. Bot. 107 (5), 723–727.

He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. 36 (4), 1–13.

Liu, S., Li, T., Chen, W., et al., 2019:. Soft rasterizer: A differentiable renderer for image-based 3d reasoning[C]//Proceedings of the IEEE/CVF. International Conference Computer Vision. 7708–7717.

Mundermann, L., MacMurchy, P., Pivovarov, J., et al., 2003. Modeling lobed leaves[C] //Proceedings Computer Graphics International. IEEE 2003, 60–65.

Ren F E, etc. 3D reconstruction of a single plant leaf image[D]. J. Image Graphics.

Tan, P., Fang, T., Xiao, J., Zhao, P., Quan, L., 2008. Single image tree modeling. ACM Trans. Graphics (TOG) 27 (5), 1–7.

Tatarchenko M, Dosovitskiy A, Brox T. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs[C] //Proceedings of the IEEE International Conference on Computer Vision. 2017: 2088-2096.

Vos, J., Evers, J.B., Buck-Sorlin, G.H., et al., 2010. Functional–structural plant modeling: a new versatile tool in crop science. J. experimental Botany 61 (8), 2101–2115.

Wang, J.L., 2013. Study on field leaf image segmentation and 3D reconstruction from a single image machine vision algorithms. China Agricultural University.

Wang N, Zhang Y, Li Z, et al. Pixel2mesh: Generating 3d mesh models from single rgb images[C] //Proceedings of the European Conference on Computer Vision (ECCV). 2018: 52-67.

D. Zermas, V. Morellas, D. Mulla, and N. Papanikolopoulos. Estimating the leaf area index of crops through the evaluation of 3d models. In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), pages 6155–6162, 2017.

Zhao, C.J., Lu, S.L., Guo, X.Y., et al., 2010. Exploration of digital plant and its technology system. Scientia Agricultura Sinica 43 (10), 2023–2030.

Zhu, L., Chen, H, et al., 2020. SingleImage 3D reconstruction algorithm based on deep learning. J. Jilin Institute Chem. Technol. 37 (01), 58–62+67.

Zhuang Y F, etc. 3D reconstruction of single image based on P2M framework[J]. Electronic Measurement Technology, 2020, 43(09): 61-64.