Phenotypic Parameters Estimation of Plants Using Deep Learning-Based 3-D Reconstruction From Single RGB Image

Genping Zhao, *Member, IEEE*, Weitao Cai, Zhuowei Wang, Heng Wu, Yeping Peng, *Member, IEEE*, and Lianglun Cheng

***Abstract*— Monitoring crop growth is of great significance to obtain crop growth status information for development of smart agriculture. The traditional way to measure the phenotypic parameters of crops is labor-intensive and encounters inconvenient operations. In this study, we propose to obtain the phenotypic parameters of crops from 3-D reconstruction of plants from single RGB images using a data-driven plant phenotypic parameters estimation network (P3ES-Net) deep neural network, which enables to estimate the depth shift and camera focal length used for depth estimation and reconstruction of the 3-D model of plants. Based on the principles of the monocular ranging and pinhole imaging model, crop phenotypic parameters such as height, canopy size, and trunk diameter can then be calculated from the 3-D model. Experiments with four practical plants present that our method is able to achieve acceptable evaluation of the growth status of plants. Of more significance, it achieves particular superior depth estimation performance over a commercial depth camera, which is a very new on-sale depth camera using stereo vision and deep learning network. This potential performance throws light on the low-cost measurement of crop phenotypic parameters using RGB camera in monitoring crop growth.**

***Index Terms*— 3-D reconstruction, deep learning, monocular depth estimation, plant phenotype parameters, smart agriculture.**

Manuscript received 29 May 2022; revised 3 August 2022; accepted 10 August 2022. Date of publication 16 August 2022; date of current version 29 August 2022. This work was supported in part by the Guangzhou Fundamental and Applied Research under Grant 202201010273; in part by the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2020B1212060069; in part by the National Natural Science Foundation of China under Grant U20A6003; in part by the Science and Technology Research in Key Areas in Foshan under Grant 2020001006832; in part by the Provincial Agricultural Science and Technology Innovation and Extension Project of Guangdong Province under Grant 2019KJ147; and in part by the Opening Foundation of Key Laboratory of Environment Change and Resources Use in Beibu Gulf, Ministry of Education, Nanning Normal University under Grant NNNU-KLOP-K1935 and Grant NNNU-KLOP-K1936. *(Corresponding authors: Zhuowei Wang; Heng Wu.)*

Genping Zhao is with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China, and also with the Key Laboratory of Environment Change and Resources Use in Beibu Gulf, Ministry of Education, Nanning Normal University, Nanning 530001, China.

Weitao Cai, Zhuowei Wang, and Lianglun Cheng are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: zwwang@gdut.edu.cn).

Heng Wu is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: hengwu@gdut.edu.cn).

Yeping Peng is with the Guangdong Key Laboratory of Electromagnetic Control and Intelligent Robots, College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Key Laboratory of Environment Change and Resources Use in Beibu Gulf, Ministry of Education, Nanning Normal University, Nanning 530001, China.

Digital Object Identifier 10.1109/LGRS.2022.3198850

I. Introduction

MONITORING crop growth has become a hotspot in smart agriculture. Being aware of the plant phenotypic parameters is one of the most fundamental tasks in crop growth monitoring [1], as it is crucial for improving the efficiency of crop production and reducing the agricultural losses. This leads the 3-D reconstruction of plant attracting great attention for acquisition of phenotypic parameters. In addition, Xie *et al.* [2] summarized the research progress of plant high-throughput phenotyping traits using unmanned aerial vehicle (UAV)-based sensors.

In recent years, 3-D reconstruction methods based on binocular vision and structured light have been widely used to monitor the growth of crops, such as the growth of rice, vegetables, and other crops. Nugroho *et al.* [3] designed a plant height monitoring system which deploys the depth-perception-based stereo camera using a parallel optical axis binocular. Yue *et al.* [4] investigated the virtual growth model of plants through establishing a human–computer interaction system based on binocular stereo vision. Peng *et al.* [5] proposed binocular camera-based 3-D reconstruction using multiple view images of the target plant. Ni *et al.* [6] made use of both monocular and binocular vision to capture image sequences of different views and used the Visual structure from motion (SFM) [7] method to generate 3-D models of plant canopies. Nguyen *et al.* [8] presented a structured light-based 3-D reconstruction system for plants to produce 3-D plant models created from multiple pairs of stereo images. Si *et al.* [9] achieved a high-fidelity 3-D plant model reconstructed based on color structured light to repair the model's depth information. Even though binocular vision and structured light-based methods render promising 3-D reconstruction in terms of accuracy, they still encounter unpleasant limitations. On one aspect, the process of binocular vision requires a large number of image sequences to get depth information through feature points' extraction and matching, which are computation-intensive. On the other aspect, the image matching accuracy greatly depends on the image textures and illumination conditions. In fact, both binocular and structured light imaging are easily affected by the ambient lighting factor. The light variations can lead to large deviations in the images and thus result in a decrease in image matching accuracy.

With the rapid development of deep learning, various deep neural networks applied for 3-D point cloud reconstruction have emerged and achieved impressive results [10], [11], [12].

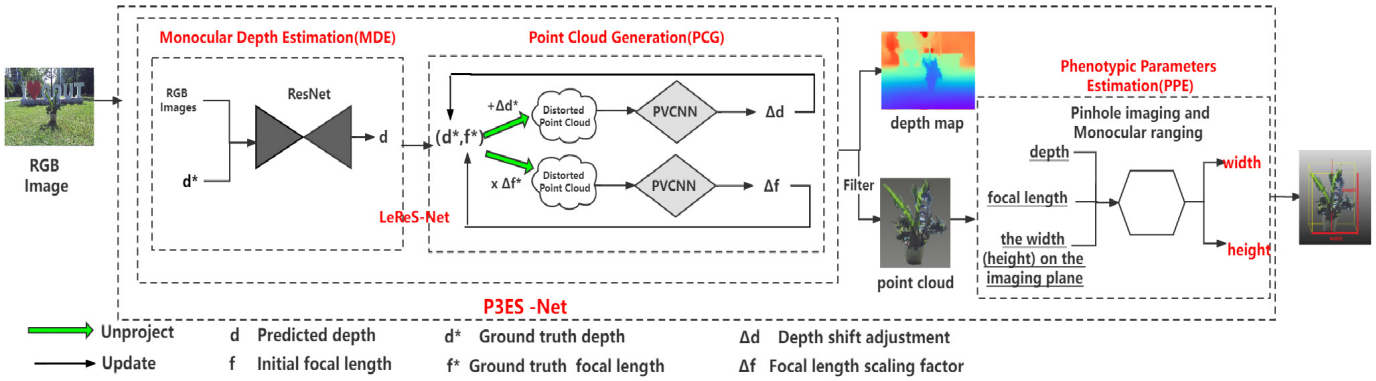

Fig. 1. Flowchart of the proposed method P3ES-Net.

Ping *et al.* [13] achieved the 3-D model of hand from a single RGB image using an I2UV-HandNet. It achieves 3-D mesh of human hand using UV-based 3-D hand shape representation; however, it cannot obtain sufficient depth information from RGB image. Pan *et al.* [14] proposed a GAN2Shape model to recover 3-D shapes from a single RGB image in an unsupervised manner which mines only some 3-D geometric cues from 2-D images generated by generative adversarial network (GAN), and depth information of the RGB image cannot be provided. Wimbauer *et al.* [15] proposed a semisupervised monocular dense reconstruction model to predict the depth map of a single moving camera in a dynamic environment. This model estimates accurate depth information from RGB images, but it is greatly affected by the environmental conditions and therefore limits its effective application under complex conditions. Yin *et al.* [16] proposed an LeReS model to estimate the depth shift and focal length for recovery of the true 3-D scene shape. This 3-D reconstruction method is the first fully data-driven method to reconstruct 3-D scene shapes from monocular RGB images. As a data-driven method, it simplifies the operational process for subsequent camera calibration in 3-D reconstruction via monocular imaging. Moreover, this method also shows superior model generalization ability, which is of great importance when applied for 3-D reconstruction of various agricultural plants. Within our best knowledge, there are few single RGB-image-based 3-D reconstruction studies in agricultural field. This is an urgent need for precise estimation of crop phenotype parameters using low-cost and simple monocular camera.

In concern of the above facts, in this study, an approach of phenotypic parameters' estimation (PPE) of plants through 3-D reconstruction from single RGB image is proposed using monocular vision imaging and the aforementioned LeReS network [16]. To evaluate its performance in real applications, a new binocular camera of ZED 2i which combines advanced depth sensing with artificial intelligence (AI) is deployed to collect plant images for 3-D reconstruction. Their 3-D reconstruction results are also used for comparison with our approach. Based on the derived 3-D models, the crop phenotypic parameters with respect to height and canopy size of the plants essential for monitoring growth conditions are estimated.

II. Method

The flowchart of the proposed approach of PPE of plant is presented in Fig. 1. First, RGB images are collected. Then preprocessing operations such as image correction including distortion correction and stereo correction are conducted on the collected RGB images. In the following, the corrected images are fed into a 3-D reconstruction deep learning framework denoted as plant phenotypic parameters estimation network (P3ES-Net). This network structure is mainly composed of two parts. One is the LeReS network [16] and the other is the PPE part following the principles of monocular ranging and pinhole imaging. As shown in Fig. 1, the monocular depth estimtion (MDE) module of LeReS initializes the depth information, and the 3-D point cloud generation (PCG) module is followed after that to estimate the missing depth shift and camera focal length so that a realistic 3-D model of the plant can be recovered. With the available 3-D model, the crop phenotypic parameters, such as height, canopy size, and other parameters, can be estimated based on the basic principles of monocular ranging and pinhole imaging.

1) LeReS Network: The LeReS network is mainly composed of two modules; one is a monocular depth prediction module, and the other is a PCG module.

The LeReS model is used to generate accurate depth map and 3-D point cloud of the target plants from singe RGB image. In the LeReS network, the depth prediction model of MDE is used to initialize monocular depth estimation. Its network structure includes a standard backbone model ResNet50 for feature extraction and a decoder for predicting depth map with unknown scale and shift. This functional part is trained in a data-driven way using multiple data sources to realize depth prediction. As these datasets have varied depth ranges, they need to be normalized to make the model training easier. To this end, the image-level normalized regression (ILNR) loss [16] is used as part of the loss function to solve this problem. It is expressed as below

$$L_{\mathrm{ILNR}} = \frac{1}{N}\sum_{i}^{N}\left|d_i - \overline{d_i^*}\right| + \left|\tanh(d_i/100) - \tanh(\overline{d_i^*}/100)\right| \tag{1}$$

where N means the number of training data samples, d_i is the predicted depth value, d_i^* is the ground-truth depth map, $\overline{d_i^*} = (d_i^* - \mu_{\mathrm{trim}})/\alpha_{\mathrm{trim}}$, and μ_{trim} and α_{trim} are the mean and the standard deviation of a trimmed depth map which has the nearest and farthest 10% of pixels removed, respectively. This function combines merits of both tanh() normalization and $Z-$ score normalization. The error of depth estimation

is simply formulated by applying a pixel-wise mean average error (MAE) between the predicted depth d_i and the normalized depth map d_i^* of ground truth.

In addition, normal is a geometric property that can effectively complement the depth information. Using the global structure as a virtual normal does not help improve the local geometric quality, such as depth edges and planes. A pair-wise normal (PWN) loss works better to constrain the global and local geometric relationships. Therefore, a PWN regression loss formulated as (2) is also included as the other part of the loss function of the depth prediction network structure to improve the performance of local geometric features' learning. PWN [16] is as follows:

$$L_{\text{PWN}} = \frac{1}{N}\sum_{i}^{N}\left|n_{Ai} * n_{Bi} - n_{Ai}^* * n_{Bi}^*\right| \tag{2}$$

where n^* indicates the ground-truth surface normal. To construct such loss function, the predicted and true depths are aligned with scale and shift factors, and the surface normal map is obtained from the reconstructed 3-D point cloud through local least-squares fitting. The planar regions with nearly identical surface normal and edges with significant changes in normal are located in the surface normal map, and points' pairs $\{(A_i, B_i), i = 0, \ldots, N\}$ whose corresponding normal are $\{(n_{Ai}, n_{Bi}), i = 0, \ldots, N\}$ are sampled globally at random on both sides of these edges. Then supervision is enforced in surface normal space to better constrain global and local geometric relationships.

The previous MDE module takes the RGB image as input and outputs an initialed depth map with unknown scale and shifts with regard to the ground truth. Then the field of view (FOV) of 60° is used to initialize the focal length and generate a distorted 3-D point cloud $\wp(u_0, v_0, f, d)$ with the predicted depth d, where (u_0, v_0) is the camera optical center. This initial point cloud is fed into the PCG module of PCG. PCG uses the Point-Voxel CNN (PVCNN) network [17] as its backbone structure to predict the depth shift adjustment Δ_d^* and focal length scaling factor a_f^*, and finally optimize the depth estimation and focal length results.

During training, the distorted 3-D point cloud $\varphi(u_0, v_0, f^*, d^* + \Delta_d^*)$ is given as the input of the PVCNN network of $N_d(*)$ to recover the depth shift using the objective function [16] as (3)

$$L_d = \min_{\theta}\left|N_d\left(\varphi\left(u_0, v_0, f^*, d^* + \Delta_d^*\right), \theta\right) - \Delta_d^*\right| \tag{3}$$

where $N_d(*)$ is the depth shift point cloud network, θ is the network weight, f^* is the ground-truth focal length, d^* is the ground-truth depth shift, and Δ_d^* is the depth shift adjustment factor. Similarly, when recovering the focal length, the distorted 3-D point cloud $\wp(u_0, v_0, a_f^*, d^*)$ is fed into the PVCNN network of $N_f(*)$ to recover the focal length using the objective function [16] (4)

$$L_f = \min_{\theta}\left|N_f\left(\wp\left(u_0, v_0, a_f^*, d^*\right), \theta\right) - a_f^*\right| \tag{4}$$

where $N_f(*)$ is the focal length point cloud network, and a_f^* is the ground-truth focal length

2) Phenotypic Parameter Estimation: To predict the growing phenotypic parameters of the measured object, the final unit (PPE) of P3ES-Net adopts the basic principles of monocular ranging and pinhole imaging. Assume d is the depth value of the measured object to the camera lens, f is the focal length of the camera lens, $G(*)$ is the practical width (height) of the measured object, and $g(*)$ is the width (height) of the measured object on the imaging plane. According to the principle of similar triangles, the practical value can be predicted following equation (5):

$$G(*) = \frac{d \cdot g(*)}{f} \tag{5}$$

where d and f can be predicted by the LeReS model, and the width (height) $g(w)$ ($g(h)$) of the measured object on the imaging plane can be calculated from (6) and (7). Then, the practical width (height) $G(w)(G(h))$ of the measured object can be achieved

$$g(w) = \frac{F_w \cdot P_w}{R_w} \tag{6}$$

$$g(h) = \frac{F_h \cdot P_h}{R_h} \tag{7}$$

where the width F_w and height F_h are related to the target area size of charge-coupled device (CCD) camera lens, P_w and P_h are, respectively, the width and height of the measured object obtained from the point cloud, and $R_w \times R_h$ is the camera resolution.

III. Experiment

A. Implementation Details

As a data-driven learning network, the MDE and PCG modules were trained using both real RGBD pairs from the public datasets of calibrated stereo DIML,[1] KITTI,[2] and synthetic RGBD pairs from the 3-D Ken Burns.[3] During training, 1000 images were withheld from all the datasets as a validation set. Those images were resized to a size of 448×448 and flipped horizontally at a probability of 0.5. The training data for each batch were loaded from different datasets evenly. The Adam optimizer was used, and a batch size of 40 and a momentum of 0.8 were set for network optimization. For the hyperparameter settings, the initial learning rate of the network was fixed at 0.02 for the MDE module and 0.24 for the PCG module, and the learning rate decayed by 0.1. The other parameters in PVCNN were set following the original work of PVCNN.

B. Experimental and Analysis

1) Data Collection: Image acquisition was conducted in the Town campus of Guangdong University of Technology (GDUT) using four practical cultivated plants with a new binocular depth camera ZED 2i for experiments.

[1] https://dimlrgbd.github.io/

[2] http://www.cvlibs.net/datasets/kitti/

[3] https://github.com/sniklaus/3d-ken-burns

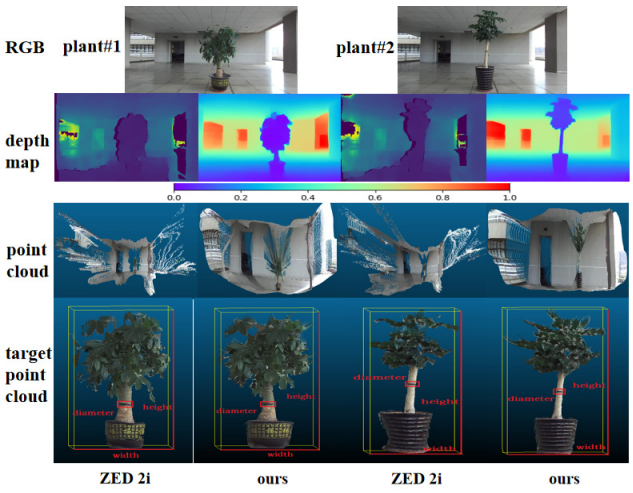

Fig. 2. 3-D reconstruction of plant#1 and plant#2 using ZED 2i depth camera and only single RGB.

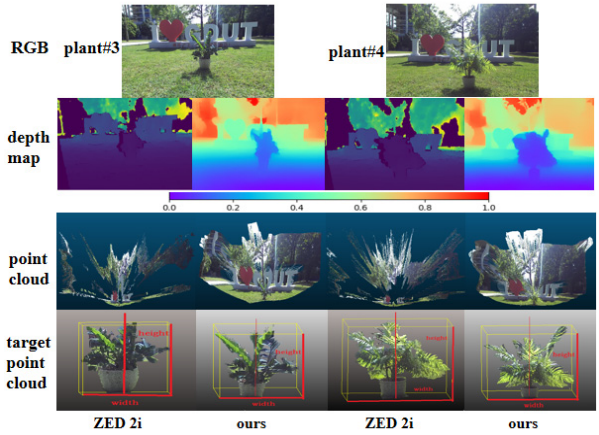

Fig. 3. 3-D reconstruction of plant#3 and plant#4 using ZED 2i depth camera and only single RGB.

ZED 2i[4] whose focal length is 2.12 mm and pixel size is 2.0 μm $\times$ 2.0 μm uses stereo vision and deep learning network for 3-D reconstruction. The camera was mounted on a tripod to collect image data from plants of plant#1, plant#2, plant#3, and plant#4 (see Figs. 2 and 3) with different backgrounds in varied outdoor environment. The first two plants Plant#1 and plant#2 with tree trunks were placed at an open balcony of the Multifunctional Building of GDUT for image collection. Plant#3 and plant#4 without tree trunk were placed in the totally outdoor environment with more complex background for imaging. For each plant, a single image was taken out from the image pair collected by ZED 2i and used for 3-D reconstruction using our method.

2) 3-D Reconstruction of Plants: Fig. 2 presents the 3-D reconstruction results of plant#1 and plant#2. It is obvious in Fig. 2 that the contour of the depth map predicted by single RGB image is much clearer than that predicted by ZED 2i, and the depth variation is more approaching the real spatial distribution. Their corresponding 3-D reconstruction results are in accordance with the depth map. The 3-D point cloud generated by single RGB image is intuitively observed with better integrity of scene, and target information which is of great significance to facilitate plant PPE is conducted on the filtered point clouds of plants as shown in Fig. 2.

Fig. 3 displays the 3-D reconstruction results of plant#3 and plant#4. The depth map obtained with ZED 2i was seen not competitive with our approach. Even though the outline of the plant targets could be somehow observed, the variations in the depth of the scene did not change naturally as human visual sensing. The depth map reconstructed from single image was of more continuously varied depth and with clear structural morphology. The corresponding 3-D reconstruction results in Fig. 3 show that under the same condition, ZED 2i captured incomplete 3-D point clouds with severe lack of scene information. On the contrary, 3-D reconstruction from single RGB image generated denser 3-D point clouds with relatively more complete scene information. Therefore, this would guarantee subsequent measurement of plant phenotypic parameters from the corresponding 3-D model which is conducted on the filtered point clouds of plants as shown in Fig. 3.

3) Phenotypic Parameters Estimation of Plants: With available 3-D models of the target plants shown in Figs. 2 and 3, PPE was conducted, and the results are summarized in Table I. The ground-truth values of the phenotypic parameters of height, canopy width, and trunk diameter were achieved using Faro Focus-S70 Lidar for much higher plant#1 and plant#2. Diameter estimation was conducted at the marked position in red as shown in Fig. 2. Manual measurement of the phenotypic parameters of height and canopy width was operated with a tape for smaller cultivated plant#3 and plant#4. Besides, different focal length estimation approaches, Zhang Zhengyou calibration method [18] and camera calibration method in MATLAB 2016, were also used to evaluate focal length estimation with our deep learning network. As can be seen from Table I, different focal length evaluation methods led to different plant phenotypic parameter evaluation results. The three parameters' estimation errors achieved for plant#1 and plant#2 using the focal length evaluation method of the LeReS network were all dominantly smaller than that using the other two methods. For 3-D reconstruction of more smaller plants of plant#3 and plant#4 in a totally outdoor environment, focal length estimation with LeReS still got competitive height estimation accuracy even though it resulted in worse parameters' estimation result with respect to the width parameter. The evidence here presents the potential use of deep learning network for focal length estimation.

Comparisons between our method and ZED 2i camera show our method got parameter estimation results on all the four cultivated plants dominantly superior to ZED 2i. When imaging taking was conducted at open balcony, our method obtained approaching but slightly better estimation accuracy with ZED 2i for width and diameter. The smallest parameter estimation error was achieved as low as 2.085% with respect to width of plant#1 by our method. As for height, the estimation errors for plant#1 and plant#2 were as high as 14.970% and 17.847%, respectively, using ZED 2i, while ours were dominantly better as 12.948% and 16.932%.

[4]https://www.stereolabs.com/zed-2i/

TABLE I

PHENOTYPIC PARAMETERS' ESTIMATION ERROR OF PLANT#1, PLANT#2, PLANT#3, AND PLANT#4

method	**focal length(left)**		**plant#1**			**plant#2**			**plant#3**		**plant#4**	
	method	value(pixel)	height	width	diameter	height	width	diameter	height	width	height	width
ground truth(mm)	——		1488.900	786.900	106.156	1773.998	782.800	54.750	628.000	582.000	898.000	884.000
ZED 2i(mm)	——		1266.007	767.847	103.033	1455.619	7588.933	53.356	483.374	437.009	726.888	807.153
errors(%)			**14.970**	**2.421**	**2.942**	**17.947**	**3.054**	**2.546**	**23.030**	**24.913**	**19.055**	**8.693**
ours(mm)	LeReS	1148.671	1681.684	803.308	103.257	2074.370	804.483	53.992	666.764	552.342	949.803	822.929
errors(%)			**12.948**	**2.085**	**2.734**	**16.932**	**2.770**	**1.384**	**6.173**	**5.096**	**5.769**	**6.909**
ours(mm)	Zhang	1073.845	1790.007	859.283	110.453	2325.113	902.659	60.581	710.170	588.299	1015.986	880.272
errors(%)			**20.224**	**9.198**	**4.043**	**31.066**	**15.312**	**10.651**	**13.085**	**1.082**	**13.139**	**0.422**
ours(mm)	Matlab	1079.159	1798.865	855.052	109.909	2313.664	898.214	60.283	706.673	585.402	1010.983	875.937
errors(%)			**20.818**	**8.661**	**3.531**	**30.421**	**14.744**	**10.106**	**12.528**	**0.585**	**12.582**	**0.912**

When ZED 2i worked in a totally outdoor environment, it provided identifiable 3-D models of plant#3 and plant#4. However, the generated 3-D models resulted in unacceptable errors in PPE when using ZED 2i. The minimum error is around 9%, while the rest were all over 19% and the highest error was even up to 25% which is far behind our data-driven method using single RGB image. The results here are in accordance with the worse qualitative results shown in Fig. 3.

IV. CONCLUSION

In this letter, we provide a single RGB-image-based deep learning 3-D reconstruction approach to acquire phenotypic parameters of crops. The approach learns from a large amount of image and depth data to learn the depth shift and camera focal lengths' parameter for 3-D model reconstruction of plants. The approach is applied on four different cultivated plants. The experiments present three-folds advantages of the approach we deployed. On one aspect, this method makes use of more extra information besides the task data to drive the deep learning network to tune and learn effective 3-D reconstruction functions. The network can be optimized with more abundant data, and the pretrained model will have good model generalization when applied to a wide variety of plants. On the other aspect, it dominantly surpassed the commercial depth camera in 3-D reconstruction tasks in outdoor environment. The estimation errors for the three parameters of plant#1 and plant#2 were estimated within 2% and 17%, respectively. They were among 3%–18% using ZED 2i. For plant#3 and plant#4, the parameter errors were testified within 5%–7%, while they were varied among 9%–25% with the advanced binocular camera in more complex background of green grass and trees. Finally, our method throws light on effective and low-cost 3-D reconstruction and measurement of crop phenotypic parameters used for smart agriculture monitoring. The single RGB image acquisition is cheap and can be easily acquired through multiple devices and platforms.

REFERENCES

[1] H. Huang, H. Zhang, C. Chen, and L. Tang, "Three-dimensional digitization of the arid land plant *Haloxylon ammodendron* using a consumer grade camera," *Ecol. Evol.*, vol. 8, no. 11, pp. 5891–5899, 2018.

[2] C. Xie and C. Yang, "A review on plant high-throughput phenotyping traits using UAV-based sensors," *Comput. Electron. Agricult.*, vol. 178, Nov. 2020, Art. no. 105731.

[3] A. P. Nugroho *et al.*, "Implementation of crop growth monitoring system based on depth perception using stereo camera in plant factory," in *Proc. IOP Conf., Earth Environ. Sci.*, 2020, vol. 542, no. 1, pp. 1–9.

[4] Y. Yue, Z. Huichun, and Z. Jiajiang, "Three dimensional reconstruction system of plant based on binocular stereo vision," *J. Chin. Agricult. Mechanism*, vol. 42, no. 3, pp. 129–135, 2021.

[5] Y. Peng, M. Yang, G. Zhao, and G. Cao, "Binocular-vision-based structure from motion for 3-D reconstruction of plants," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3105106.

[6] Z. Ni, T. Burks, and W. Lee, "3D reconstruction of plant/tree canopy using monocular and binocular vision," *J. Imag.*, vol. 2, no. 4, p. 28, Sep. 2016.

[7] C. Wu. (2011). *VisuaLSFM: A Visual Structure From Motion System*. [Online]. Available: http://www.cs.washington.edu/homes/ccwu/vsfm

[8] T. T. Nguyen, D. C. Slaughter, N. Max, J. N. Maloof, and N. Sinha, "Structured light-based 3D reconstruction system for plants," *Sensors*, vol. 15, no. 8, pp. 18587–18612, 2015.

[9] K. Si, J. Zhang, Z. Li, Z. Guo, X. Lu, and J. Xie, "High-fidelity 3D plants model reconstructed based on color structured light," in *Proc. 3rd Int. Conf. Agro-Geoinform.*, Aug. 2014, pp. 1–4, doi: 10.1109/Agro-Geoinformatics.2014.6910587.

[10] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 945–953.

[11] Y. Yao *et al.*, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.

[12] S. Im, H.-G. Jeon, S. Lin, and I. S. Kweon, "DPSNet: End-to-end deep plane sweep stereo," 2019, *arXiv:1905.00538*.

[13] C. Ping *et al.*, "I2UV-HandNet: Image-to-UV prediction network for accurate and high-fidelity 3D hand mesh modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 12929–12938.

[14] X. Pan, B. Dai, Z. Liu, C. C. Loy, and P. Luo, "Do 2D GANs know 3D shape? Unsupervised 3D shape reconstruction from 2D image GANs," in *Proc. ICLR*, 2021, pp. 1–18.

[15] F. Wimbauer, N. Yang, L. Von Stumberg, N. Zeller, and D. Cremers, "MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera," in *Proc. CVPR*, 2021, pp. 6112–6122.

[16] W. Yin *et al.*, "Learning to recover 3D scene shape from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 204–213.

[17] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel CNN for efficient 3D deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 3, no. 5, pp. 1–11.

[18] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.