Simultaneous Direct Depth Estimation and Synthesis Stereo for Single Image Plant Root Reconstruction

Yawen Lu, *Student Member, IEEE*, Yuxing Wang, Devarth Parikh, Awais Khan, and Guoyu Lu, *Member, IEEE*

***Abstract*— Plant roots are the main conduit to its interaction with the physical and biological environment. A 3D root system architecture can provide fundamental and applied knowledge of a plant's ability to thrive, but the construction of 3D structures for thin and complicated plant roots is challenging. Existing methods such as structure-from-motion and shape-from-silhouette require multiple images, as input, under a complicated optimization process, which is usually not convenient in fieldwork. Little effort has been put into investigating the applications of deep neural network methods to reconstruct thin objects, like plant root systems, from a single image. We propose an unsupervised learning scheme to estimate the root depth from only one image as input, which is further applied to reconstruct the complete root system. The boundaries of the reconstructed object usually contain large errors, which is a significant problem for roots with many thin branches. To reduce reconstruction errors, we integrate a cross-view GAN-based network into the reconstruction process, which predicts the root image from a different perspective. Based on the predicted view, we reconstruct the root system using stereo reconstruction, which helps to identify the accurately reconstructed points by enforcing their consistency. The results on both the real plant root dataset and the synthetic dataset demonstrate the effectiveness of the proposed algorithm compared with state-of-the-art single image 3D reconstruction models on plant roots.**

***Index Terms*— Root reconstruction, cross-view synthesis, single image depth estimation.**

I. Introduction

ROOTS are the main conduits that are intermingled with the physical and biological environment. Root system influences resource uptake from the soil and is a key for plants to thrive [1], [2]. The root system architecture of plants is complex, with various branches and shapes but fairly uniform in color, which makes it difficult to reliably extract features. However, single-image 3D scene reconstruction is a long-pursued goal in computer vision. A detailed and complete reconstructed 3D structure can help in subsequent 3D tasks like detection, segmentation, and recognition. 3D root models can provide key features needed for plant breeders and biologists to measure various traits including root length, volumetric biomass and crop growth. Recently, significant progress has been made in contributing synthetic datasets for shape reconstruction [3]–[6] and generating 3D shapes from a single image by using deep neural networks [7]–[11]. However, the above-mentioned measures heavily rely on estimating the 3D shapes by learning geometric representation from ground truth 3D models or CAD priors. In practice, most of the ground truth 3D models for objects are not available. Also, it is not always feasible to scan the surface of the objects with professional quality equipment. Specifically, for plant root reconstruction, existing 3D reconstruction methods may recover the basic structure of the root system. However, these methods, generally, fail to reconstruct in-depth and fine regions of the complex root structures, which have few textures for feature extraction and matching. Recent reconstruction methods either present low point cloud density, which leads to incomplete models with occluded regions and large holes [12], or require multiple cameras to capture many images at the same time [13], [14], which increases the reconstruction cost.

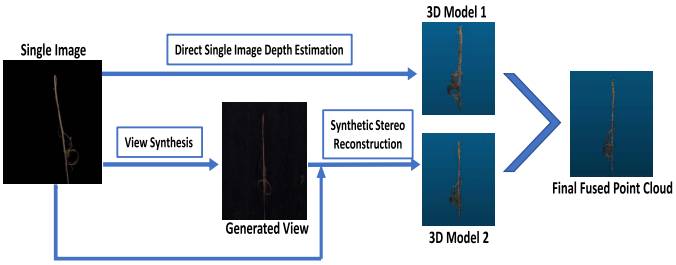

Fig. 1. During inference, the pipeline is able to generate a fused point cloud from the direct single image depth estimation and the synthesis stereo reconstruction to construct more accurate and complete 3D models for plant roots, given one single image as input.

In this paper, we aim to reconstruct the roots simply from a single image that anyone would find easily usable, even during remote fieldwork or in the laboratory or greenhouse, as shown in Fig. 1. As plant root shapes across species vary significantly (e.g., potato vs grass root vs tree roots), we propose a novel framework to reconstruct the 3D plant roots, in an unsupervised manner, by using one single-view image that can learn the system without labeling issues. Furthermore, our model learns a cross-view synthesis model

Manuscript received August 16, 2020; revised January 25, 2021 and March 7, 2021; accepted March 13, 2021. Date of publication April 20, 2021; date of current version May 12, 2021. This work was supported by the Engineering for Agricultural Systems program through the United States Department of Agriculture (USDA) National Institute of Food and Agriculture under Grant 2021-67021-34199. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yi Yang. *(Corresponding author: Guoyu Lu.)*

Yawen Lu, Yuxing Wang, Devarth Parikh, and Guoyu Lu are with the Intelligent Vision and Sensing Laboratory, Rochester Institute of Technology, Rochester, NY 14623 USA (e-mail: luguoyu@cis.rit.edu).

Awais Khan is with the Plant Pathology and Plant-Microbe Biology Section, Cornell University, Ithaca, NY 14850 USA (e-mail: awais.khan@cornell.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2021.3069578, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3069578

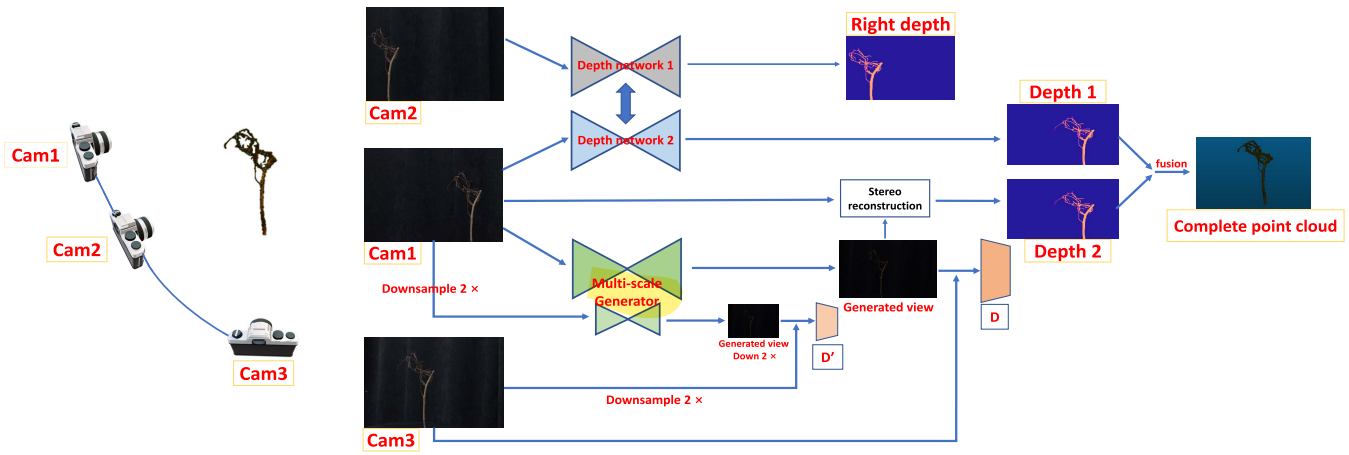

Fig. 2. Overview of the proposed plant root reconstruction training pipeline. The proposed method takes advantage of view synthesis-based synthetic stereo reconstruction and direct single image depth estimation to output accurate and stable 3D point clouds. Although trained with images from different perspectives, the proposed method requires just one single-view image as input for inference.

to simulate a different view, since partial observation from a single perspective would cause large errors in the reconstructed 3D model for unseen regions. The predicted picture can reconstruct the root, together with the input image, by stereo setting. Finally, the two corresponding point clouds from the two different 3D reconstruction approaches can constrain each other to recover a combined and accurate 3D structure. The generated 3D shape from a single input image can prevent a large portion of outliers compared with other state-of-the-art models. To evaluate the performance of the proposed single image root reconstruction method, we have conducted experiments on a newly collected dataset for roots of one-year-old apple trees and a synthetic dataset, with comparisons to other state-of-the-art 3D reconstruction methods. The main contributions of this work are listed below:

1) We introduce a self-supervised deep neural network to directly generate a 3D model from a single RGB image, benefiting from the introduced Atrous Spatial Pyramid Pooling (ASPP) module in the modified depth estimation network structure and the edge and keypoint consistency constraints specially designed for thin roots; 2) A synthetic stereo reconstruction approach is developed based on the multi-scale discriminator or generator structures on neighboring-view prediction to learn both holistic and local features; 3) A fusion strategy is proposed to fuse the 3D point clouds from two different single image depth estimation methods, to eliminate unexpected outliers and to generate an accurate and complete refined point cloud.

Unlike existing learning-based methods, the proposed framework is trained in a self-supervised manner. The entire training pipeline is shown in Fig. 2. Cam 1, 2 and 3 denote three cameras and their corresponding captured images. Camera 1 and camera 2 are combined to train the direct depth estimation, and camera 1 and camera 3 contribute to the synthetic stereo reconstruction. During inference, only a single image from camera 1 is utilized as input to reconstruct a refined 3D root model.

II. Related Work

Laser scanning and structured light methods have been widely used in 3D reconstruction for their high accuracy and sensitivity. However, due to their limitations, such as high price and poor portability, active acquisition techniques are not feasible. Recently, image-based passive techniques have been explored to act as a substitute of active methods such as multi-view stereo [15]–[17], SfM [18], [19], single image [7]–[10] and image sequences [20]–[23].

Multi-View Reconstruction: Methods attempt to estimate and refine the depth maps. Reconstructed dense depth maps from the sparse point cloud can be utilized to remove unwanted points that are significant in conflict with each other [15]. MVSNet [16] proposed to utilize a deep neural network architecture for depth map estimation, which boosts the completeness of the reconstruction model and its overall quality. MVDepthNet [17] designed a convolutional neural network to tackle the multi-view depth estimation problem given several image-pose pairs from a localized monocular camera with neighboring viewpoints. Structure-from-motion (SfM) based reconstruction is a process of reconstructing 3D structures from multiple images taken from different viewpoints. With a large number of unordered images, SfM [18], [19] incrementally and sequentially reconstructs the scene with an iterative component by continuously registering new images, then triangulating the 3D points, and refining the reconstruction results with RANSAC [24] and bundle adjustment [25]. [19] is a global optimization approach that attempts to recover a consistent viewing graph. However, those methods require multiple image inputs and sophisticated optimization methods, which greatly increases processing complexity and memory consumption.

Supervised Learning-Based Reconstruction: Generates 3D models from a single image [7]–[10] based on the development of deep learning techniques. Choy *et al.* [8] applied convolutional neural networks to learn a mapping from observations to their underlying 3D shapes of objects from a large collection of training data with ground truth labels.

[7] explored generative networks for 3D model generation based on a point cloud representation instead of mesh or CAD model. [9] proposed to utilize generative adversarial networks to generate the complete 3D occupancy grids by taking the voxel grid representation of the depth of the input object. However, the aforementioned data-driven algorithms heavily depend on the ground truth 3D model for training, which is not always available and feasible to be collected in real applications.

Unsupervised Single Image Depth Estimation: Takes a single image as input for depth estimation during testing [20]–[22], with the target to deal with single-view depth prediction together with camera pose regression using unlabeled video sequences based on photometric consistency between source and target view. Godard *et al.*, aims to predict the pixel disparities from synchronized stereo pairs [26]. However, the methods mentioned above mainly target outdoor scenes, which contain extensive textures for matching instead of thin and sophisticated objects like plant roots. Also, the reconstruction model from one view would have a large portion of infinite points due to the lack of multiple view constraints, which may cause large errors.

Cross-View Synthesis: Transforms the information of the same objects such as texture, shape, and color of input view to synthesize different views [27], [28], such as a standard encoder-decoder network to obtain the 3D feature representation of the training data and then predict objects' 2D images from different views during testing [27]. Alternatively, predicting the target view from the source can be avoided. Appearance flow indicated that pixels can be predicted in the input view to reconstruct the target view [28]. With the success of Generative Adversarial Networks [29], conditional GANs have been used to synthesize cross-view images [30]–[32]. Among them, [31] translated text from human-written descriptions to image synthesis of the scene. This is further refined by deploying two-stage stacked GAN networks to generate high-resolution images with photo-realistic details [32]. Zhai *et al.* [30] proposed to transfer dense pixel-level labels of the ground imagery to aerial imagery. They also utilized the predicted images to produce their corresponding ground-level panorama.

The cycle-based synthesis network learns a geometric mapping from one view to another, which does not require query images and is directly simulated from source view input. Compared with the aforementioned expensive root reconstruction solutions that suffer from the fine root branch occlusion and difficulty to extract surface feature and color, we propose to utilize a single image root reconstruction method to recover the original root structure from both spatial and spectral perspectives. In this paper, We first develop two different image depth estimation methods, direct single image depth estimation and synthetic stereo reconstruction based on view synthesis, that simultaneously estimate the image depths. Then we fuse the two reconstructed models based on their spatial consistency. The final reconstructed model takes advantage of both methods to generate an accurate and complete 3D point cloud.

III. Simultaneous Single Image Depth Estimation

The single image 3D reconstruction method involves two simultaneous single image depth estimation methods with different benefits. One is the direct single image depth estimation approach that explores the end-to-end unsupervised single image depth estimation based on left and right training images for thin and complicated plant roots. The other is to generate a synthesized cross-view image for synthetic stereo reconstruction. The direct single image depth method can have a better front-facing view reconstruction due to its strict disparity constraints. However, as it misses the side view, the boundary region's depth could have large errors. The synthesis view can help constrain the large boundary error from a different perspective during the synthetic stereo reconstruction. However, due to the large view changes, some front-facing regions have relatively more errors. Therefore, we apply both depth map outputs and take advantage of each method. Based on the two simultaneous depth outputs, the estimated depth maps from both outputs constrain each other to generate a refined 3D shape model for the single root image.

A. Direct Single Image Depth Estimation

In order to get the depth map of the thin and complicated plant root structure, we conduct the single image depth estimation from real stereo image pairs during the training process. The image depth estimation problem is regarded as a 2D image reconstruction task by projecting the image from the left view to the right view with the help of the right disparity map and minimize the difference of the reconstructed right image and original right image, and vice versa. We adopted the coarse-to-fine learning strategy and applied the residual pyramid that allowed learning and refinement of disparities in one decoder. Similar to [26] and [33], we adopted ResNet-18 [34] as the encoder to extract features and local information, and the corresponding decoder with skip connections to recover higher resolution details. Different from the two aforementioned works, we added an Atrous Spatial Pyramid Pooling module (ASPP) with a dilation rate of 1, 6, 12 and 18 to the end of the encoder network to better learn contextual information, inspired by the promising performance of this module in DeepLab v3 [35] on segmentation tasks. Instead of generating two disparities of the left view and right view given only the left view as input as in [26], [33], our model equally treats the left and right views, outputs one disparity for each of them, and associates the two CNNs with weight sharing, as depicted in Fig. 3.

Depth estimation is initialized on a small scale and further recovered on a large scale with more high-frequency representations and details. The corresponding point cloud is then obtained from the predicted disparity map, preset baseline and focal length set for training after calibration.

More specifically, given each image pair from the stereo setting, the proposed network builds two CNN networks to predict their disparity maps. Following this concept, the predicted left and right disparity maps are applied to reconstruct the original left and right input image, respectively. To keep

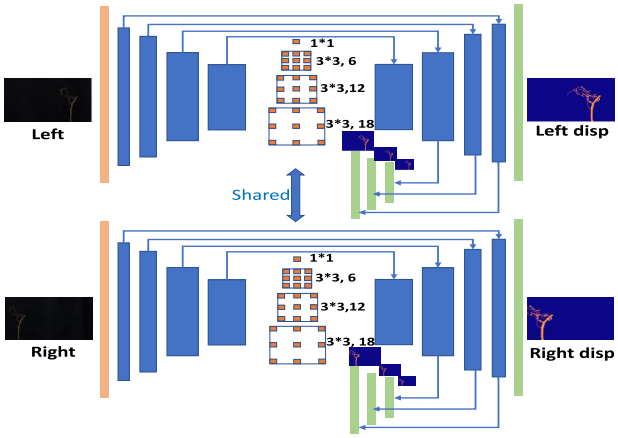

Fig. 3. Network architecture of the direct single image depth estimation method.

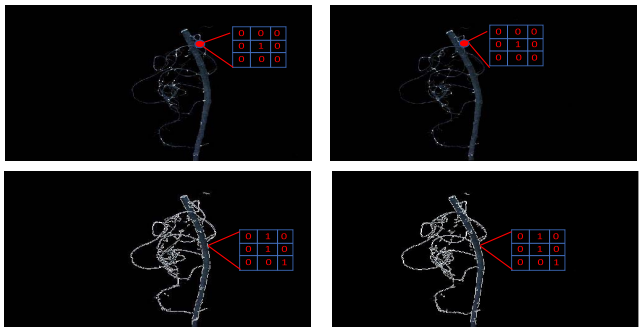

Fig. 4. Designed spatial loss constraints for edge maps and keypoints. This loss term applies hamming distance to constrain the predicted disparity map matching the left and right edge maps and keypoints at each pyramid scale.

the consistency between the reconstructed images and the original input image, and minimize the difference between them, we introduce the spectral consistency L_{consis} loss term, which combines L1 loss, L2 loss, and SSIM term to penalize the possible difference in the predicted regions:

$$L_{consis} = \frac{1}{N}\sum_{ij}\frac{a}{2}(1 - SSIM(I_{ij}, Warp(I_{ij})) + \frac{2}{3}(1-a)(||I_{ij} - Warp(I_{ij})||_1) + \frac{1}{3}(1-a)(||I_{ij} - Warp(I_{ij})||_2) \quad (1)$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ represent L_1 and L_2 norm operators to calculate mean absolute and Euclidean distance. I_{ij} represents the original inputs (either left or right image), and $Warp(I_{ij})$ represents the warped image with disparity map by moving pixels from the original input image along the epipolar line. a is chosen to be 0.85 based on testing.

Unlike outdoor street scenes and indoor scenes, the texture and color of plant roots are rather similar, which could result in wrong matching. We propose to constrain the disparity map further to maintain the consistency of Canny edge maps and SIFT keypoints between the left and right images, which is more effective than conventional methods just considering the variation of raw pixel intensities, as shown in Fig. 4.

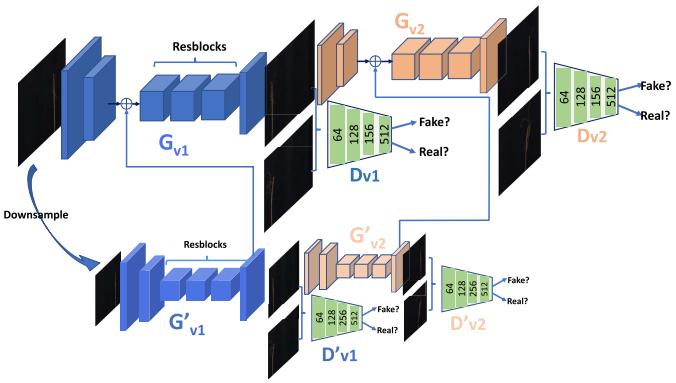

Fig. 5. Network architecture of the cross-view synthesis framework.

For keypoints here, the 50 strongest SIFT feature points on the generated edge mage are extracted. Therefore, we can further rely on spatial consistency to optimize the disparity.

The consistency term for keypoints matching using the horizontal positions of paired keypoints in the left and right images $L_{keypoints}$ is defined as:

$$L_{keypoints} = \sum_i \sum_{j \in N_i} \left\| (KP_i - KP_j) - (\hat{KP}_i - \hat{KP}_j) \right\|^2 \quad (2)$$

where N is a set of keypoints KP_i detected in the left and right images. $KP_i - KP_j$ is the ideal disparity difference, and $\hat{KP}_i - \hat{KP}_j$ is the actual difference for the estimated disparity. And the consistency term for edge map is defined as:

$$L_{edge} = \sum_{ij} Hamm\_Dist((||E_{ij} - Warp(E_{ij})||)) \quad (3)$$

where E_{ij} is the detected edge map. $Hamm\_Dist$ represents the Hamming distance between the two image windows. Thus, the completed loss constraint for the edge map and matched keypoints becomes:

$$L_{key\_edge} = \sum \alpha \times L_{edge} + (1-\alpha) L_{keypoints} \quad (4)$$

where α is set to be 0.7 and thus $1-\alpha$ to be 0.3 to balance the two consistency terms.

B. Synthetic Stereo Reconstruction Based on View Synthesis

Simultaneously, we reconstruct the plant root 3D model based on synthetic stereo reconstruction, which requires a synthesis view from a different perspective. Different from [36] using supervised CNN to output synthesis images, our method relies on a newly developed GAN structure targeting at root images to generate the synthesis image better, fitting both root thin and complicated branch properties. Extending [37] which used residual blocks to learn the translation function, we further apply the multi-scale generator and discriminator structure, as shown in Fig. 5. The increased number of discriminators and generators are able to enforce the model to learn more detailed information. The multi-scale generators consist of local (G_{v1}, G_{v2}) and global generators (G'_{v1}, G'_{v2}). Each includes a convolutional front-end, three residual blocks and a deconvolution back-end. The input to G'_{v1} and G'_{v2} is the

twice down-sampled image of the raw input. The feature maps from the output of G'_{v1} and G'_{v2} are element-wise summarized with the features output from the front-end of G_{v1} and G_{v2}, as the new input to the residual blocks of G_{v1} and G_{v2}. By this way, the features are extracted at both coarse and fine scales. The discriminators (D_{vi}, D'_{vi}) are to distinguish whether the full-size and down-sampled images are real or fake, with each composed of five convolutional layers with 64, 128, 256 and 512 filters, respectively.

The adversarial loss here is for the discriminator to distinguish the real images from the generated images. Here, we represent the adversarial loss for the network as L_{Adv}:

$$L_{Adv,v1\to v2} = \sum_{l=1}^{L} E_{t\in p_{v1(c)}}[1 - log(D(c)^{(l)})] + E_{t\in p_{v2(t)}}[logD(G(t)^{(l)})] \quad (5)$$

where the adversarial loss $L_{Adv,v1\to v2}$ is used to synthesize and learn a mapping function from the source view $v1$ to target view $v2$. The discriminator D is used to judge whether the generated image is the same as the original image or not. Once the input is the image from source view $v1$ in dataset c, the discriminator can reduce the loss by outputting 1 from $[1 - logD(c)]$. Similarly, if the input is an image from the target view $v2$ in the original dataset t, the discriminator is expected to output 0 for the generated image $G(t)$. Multi-scale discriminator $D^{(l)}$ and generator $G^{(l)}$ are utilized to produce more accurate performance.

In cross-view synthesis, cycle consistency is applied as the structural constraint of the cycle-based framework. The synthesis image from the generator can be the input of the auxiliary generator to output a simulated image from the source view. The cycle consistency is to enforce the input image and the cross-view predicted images as close as possible. The loss constraint is denoted as the following equation:

$$L_{cycle} = \sum_{l=1}^{L} E_{c\in Ori(c)}[\left\| G_{v2}^{(l)}(G_{v1}^{(l)}(c)) - c \right\|_1] \quad (6)$$

The cycle loss L_{cycle} is introduced to further reduce the possible difference of the training samples distance between generated view image $G_{v2}(G_{v1}(c))$ and the original view image c. The generator G_{v1} translates the original view images c to a cross-view image $G_{v1}(c)$, which is further translated to the original view image $G_{v2}(G_{v1}(c))$ by the auxiliary generator G_{v1}. Through the cycle consistency, though both networks are trained in random, better image quality and stability can be achieved compared with the structure without it. Hence, the full objective functions that optimize the network parameters becomes $L_{total} = \lambda_{consis}L_{consis} + \lambda_{key\_edge}L_{key\_edge} + \lambda_{adv}L_{adv} + \lambda_{cycle}L_{cycle}$. λ_{consis}, λ_{key_edge}, λ_{adv} and λ_{cycle} represent the weights for different loss constraints and are set as 1, 0.2, 0.2, 1, 10 respectively.

After simulating the cross-view, the source view and the predicted target view are used to compute the disparity map using stereo matching and hence compute the depth map. Also, we use the predicted disparity map for the source view to compute another depth map. The depth map is further utilized to project the image pixels into 3D space as:

$$Z_i = \frac{f B_i}{x^l - x^r}, \quad X_i = \frac{(x^l - c_x)Z_i}{f}, \quad Y_i = \frac{(y^l - c_y)Z_i}{f} \quad (7)$$

where Z_i and B_i are the depth and baseline for each method, respectively. f is the focal length. (c_x, c_y) is the principal point of the image. (x^l,x^r) are the horizontal coordinates and (X_i, Y_i, Z_i) represents the 3D point's coordinate. By comparing the Euclidean distance d_{3D} of these two point clouds in 3D space, we merge those points that d_{3D} is smaller than the threshold T, and exclude those points that d_{3D} is larger than the threshold to generate the final completed 3D model. By removing the inconsistent points from the two point clouds and exploiting their geometric consistency information, we leverage advantages from both the synthetic stereo reconstruction method and the direct depth estimation method and are capable of getting the fused point cloud which is more detailed and accurate in 3D representation of the roots. From visual experiments, the proposed method can get rid of most outlier noise and infinite points that commonly happen in single image depth estimation algorithms.

TABLE I

THE THREE CAMERAS (CANON EOS REBEL SL1) ARE OF SAME MODEL AND SHARE THE SAME SETTING DURING DATA COLLECTION

Sensor Name	Canon EOS SL1
Sensor Type	CCD
Pixel Dimension	2592 × 1728
Length of Focus	20 mm
Sensor Size	22.3 × 14.9mm
FOV	58 ×41

IV. EXPERIMENTS

In this section, we first describe the approach used to collect dataset. We then introduce the configuration and the evaluation of the proposed model on the collected dataset, and present the training and evaluation setup. Later, both qualitative and quantitative results are compared with other recent methods. We show that the performance of our model without using ground truth 3D models for training is superior compared to other methods. In the end, an additional experiment is done to verify the generalization of the proposed method on public scenes in real and synthetic datasets.

A. Dataset Introduction

In order to tackle the target problem in this work, a high-resolution real-world dataset captured by three Canon EOS DSLR cameras, namely the "DSLR Root Dataset" $(DR)^2$ is introduced, to generate a complete 3D root model from single image as input during inference. The parameters of devices used to collect the data are described in Table. I and examples of the collected plant root dataset for apple trees and synthetic root images are shown in Fig. 6. The collected dataset has high-resolution RGB images of 23 apple root system captured

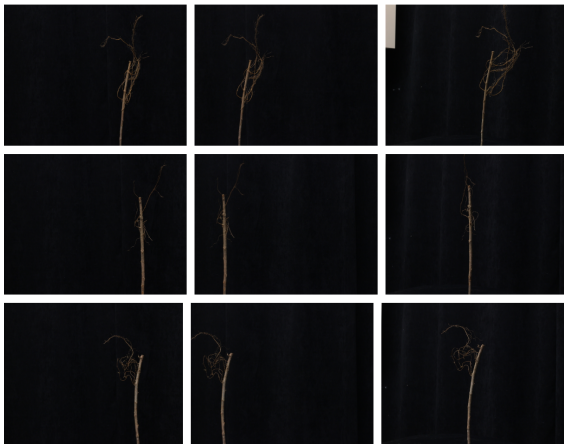

Fig. 6. The collected root dataset with each row of images captured from one of the 3 different viewpoints. Note that root system images are from one year-old apple rootstock cuttings obtained from a commercial nursery and are shown upside-down.

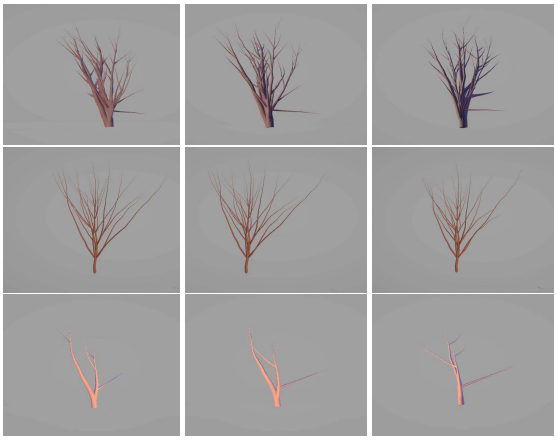

Fig. 7. Synthetic root dataset with each row of images captured from one of the 3 different viewpoints.

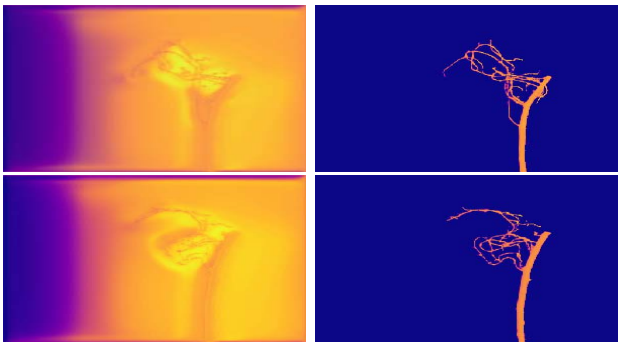

Fig. 8. Direct depth estimation results on given examples without (first column) and with (second column) segmentation on the raw images.

by three DSLR cameras with the same setting. These apple root system used to capture images are from one year old apple rootstocks generated from cuttings by a commercial nursery. The images of root systems of apple rootstocks were taken/shown upside-down. We take the center camera 1 as the reference camera (world coordinate origin) and first calibrate camera 1 and camera 2 to obtain intrinsic and the extrinsic between them, followed by calibrating camera 1 and camera 3. The collected images from the three cameras can then be rectified and described by the same world coordinate. Camera 1 and camera 2 in Fig. 2 for direct depth estimation network training are positioned with mainly a translation motion and the least rotation. Camera 1 and camera 3 are positioned with an angle of approximately 60 degrees. We rotate each root at a constant speed to capture 150 images covering views from 360 degrees. We split the dataset into 18 roots' images for training and 5 roots' images for testing. In total, we have 2700 images for training and 750 images for testing. To ensure all cameras to capture photos simultaneously, they were mounted on a tripod and activated by a wireless remote controller. The ground truth 3D root models come from a 3D active laser scanner.

In addition to the real dataset, we also create a synthetic dataset based on the GameObject class in Unity game engine [40]. By creating a synthetic 3D model for each plant root, and controlling the 3D spatial relations of three virtual cameras to be the same as in the collected real dataset, we enjoy the same setting and training/test data size as the real dataset. Sample images from our created synthetic dataset are shown in Fig. 7.

B. Training Configuration

For cross-view prediction, we first scale the images to 720×480. During each iteration, one source view (Cam 1) and one reference view (Cam 3) were used to train the view prediction network. After training, the source view and predicted view are then applied for stereo reconstruction to predict a multi-view depth map. The source image is also used for direct depth estimation. We later utilize the depth maps to produce root 3D point clouds. The framework is implemented with PyTorch and trained from scratch using Adam optimizer [41] with $\beta 1 = 0.9$ and $\beta 2 = 0.99$. The initial learning rate starts with 1e-4 and gradually decays to half for every 10 epochs in our framework. The weights of the depth estimation and view synthesis network are initialized with Kaiming initialization [42] with a batch size of 4. The framework is trained for 30 epochs.

C. 3D Root Reconstruction

We use the test split to evaluate our models based on the following metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) to evaluate the performance of view synthesis; Fitness score and Mean distance to evaluate the performance of the reconstructed 3D model. PSNR is defined as $PSNR = 10log_{10}(\frac{MAX^2}{MSE})$, where MAX is the maximum pixel value and MSE is the Mean Squared Error between the pixel values in the original image and the simulated output. The Structural Similarity Index (SSIM) is a perceptual metric that qualitatively measures the similarity between the source and target image. It is expressed as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (8)$$

Fig. 9. Visual performance of our view synthesis method. First column: Single image input; Second: Cross-view CGAN [38] output; Third: SingleGAN [39] output; Fourth: Our method output (with more details and less blur); Last column: Ground truth image of the target view.

where μ_x, μ_y, σ_x, σ_y, and σ_{xy} are the local means, standard deviations, and cross-covariance for source and target images, respectively. The mean distance is the mean distance value of each point relatively to the reference ground truth point cloud, and the variance reflects the deviation of the generated point cloud with the ground truth.

To avoid the effect from the background, we first segment the background, as background can generate a disparity different from the root foreground, especially when background intensity between images is close. The segmentation effect for disparity estimation is shown in Fig. 8. It can be observed that with the background segmentation, the disparity can be much more accurate compared with the result without segmentation.

D. Comparison With State-of-the-Art Methods

We first provide a qualitative comparison against state-of-the-art methods on view synthesis and 3D reconstruction. We then report the numerical comparison and analysis. To make a fair comparison, all compared methods are re-trained on the same data splits. For cross-view synthesis, we directly use the images from Cam1 (source view) and Cam3 (reference view) to train all other methods [30], [39], [45]–[47]. For 3D reconstruction, we feed left-view images from Cam1 and right-view images from Cam2 to train multi-view depth estimation algorithms [17], [26], [33], [43], and left-view images from Cam1 to train the single-view depth estimation algorithm [44].

1) Qualitative Comparisons: For qualitative results, our method is visually evaluated for predicting the view with the most recent methods: Pix2Pix [45] and SingleGAN [39]. Our approach can predict the image from a different perspective, provided a single input image with a known viewpoint. Fig. 9 shows the generated synthesized images with adversarial and cycle-consistency losses, which demonstrates that our method is effective in simulating other views for complicated and thin objects like plant roots. Compared with other methods, Pix2Pix [45] preserves the main structure of the branch. However, it is not effective for predicting detailed objects. In addition, their synthetic images contain many blurred regions. SingleGAN [39] method can recover more detailed shapes and branches, but most of the predictions have a considerable difference in intensity compared with the original images.

We also compare with other 3D reconstruction approaches retrained on our collected dataset, as Fig. 10. The results demonstrate that our proposed model is able to generate a more accurate point cloud from a single image reliably. Many stretching and wrong coloring artifacts and infinite black pixels on [26], [43] and [44] can be observed because they are unable to observe the back regions of roots from just a single image, from one fixed viewpoint. Without many stretching regions and black pixels, [33] and [17] suffer from severe deformation in shape and wrong-prediction in object depth. With the help of view synthesis network and point cloud fusion, we are able to restrict the point cloud more tightly and exclude outliers in an unsupervised way. These comparisons demonstrate the superiority of our framework in effectively eliminating the incorrect depths and generating the accurate 3D point cloud.

Our method for direct depth estimation takes stereo images as inputs for training and only one single image for testing in the real applications. Compared with our method, [17] takes multiple images for training and two images at different views to estimate the inverse depths. As shown in Fig. 13, compared with our method, images simulated and generated with [17] suffer from intense discontinuity on thin branches of roots.

We perform a series of ablation studies to analyze the benefits of our proposed components, in Table II, Table III, Fig. 11, Fig. 15 and Table IV. Visual comparisons on the produced 3D point clouds are demonstrated in Fig. 11 and Fig. 15, and quantitative comparisons are provided in Table II, III and IV, compared with only applying direct single image depth estimation or synthetic stereo method. It can be observed that although our direct single image depth estimation and synthetic stereo methods can both generate relatively accurate point clouds, they still have errors in different perspectives, which can be improved by enforcing their spatial consistency to fuse the reconstructions.

To compare with other approaches on stronger ground truth models, we further conduct a comparison on the created synthetic dataset, as shown in Fig. 12. We notice a visual

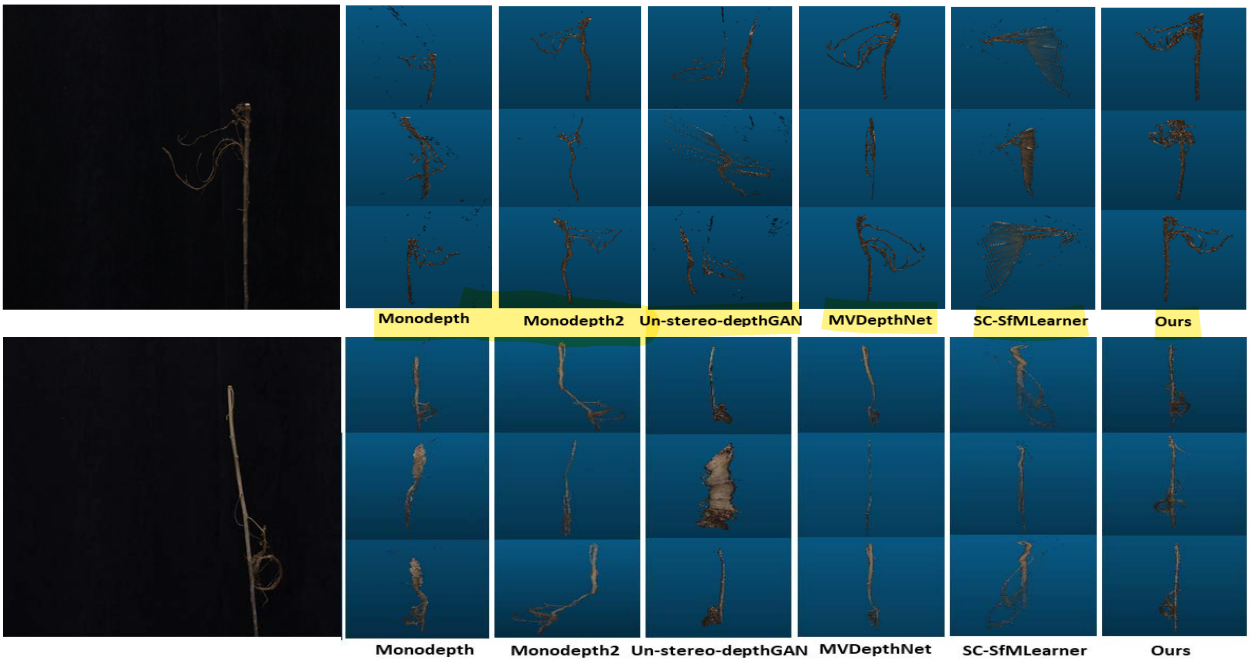

Fig. 10. Comparison of reconstructed roots between our method and other state-of-the-art methods. From left to right: Input, Monodepth [26], Monodepth2 [33], Un-stereo-depthGAN [43], MVDepthNet [17], SC-SfMLearner [44], and our proposed method. Our reconstruction model can capture more accurate shape of root structure.

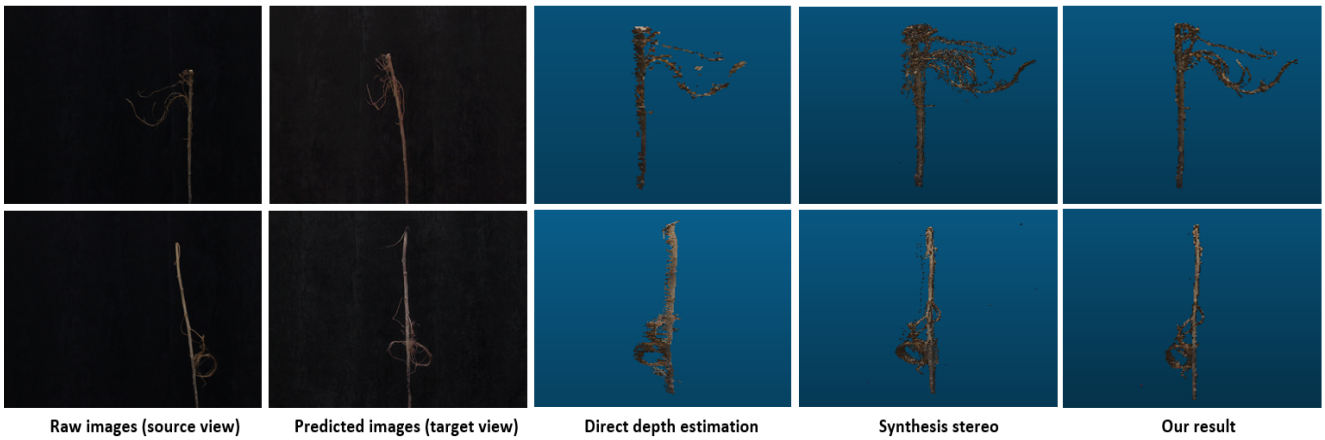

Fig. 11. Visual performance of our proposed method on generating final complete point cloud for roots compared with only from synthesis stereo reconstruction and direct depth estimation method. Compared with only applying the synthesis stereo reconstruction and direct depth estimation method, the fused model can capture more details in plant root structures and also prevent some blurring and wrong-matching regions.

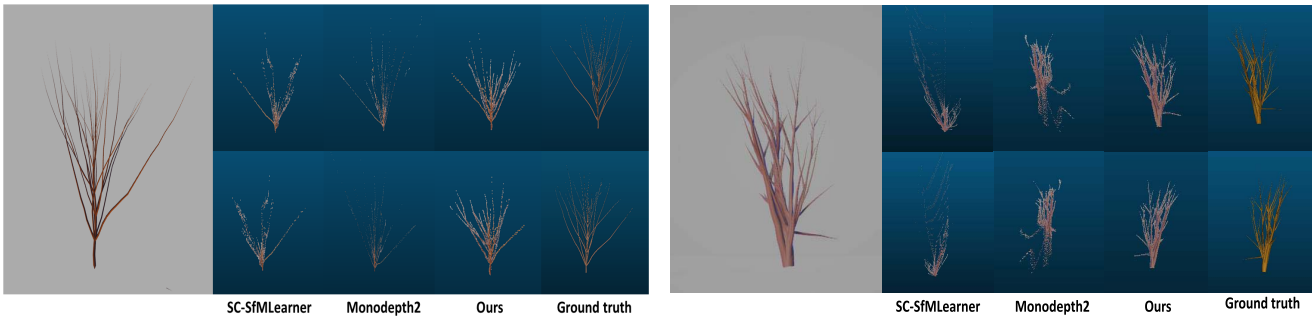

Fig. 12. Visual performance of our proposed method compared with other methods on synthetic data. Compared with other methods on the synthesis dataset, the proposed method can reconstruct much more branches as the ground truth.

improvement in distortion and stretching issues from Fig. 12 for most methods, which could be explained from the high-quality rendered images of the created 3D model with a clear background. However, a great number of thinner root branches also result in an even more challenging situation compared with real collected data. Compared with the results from [33],

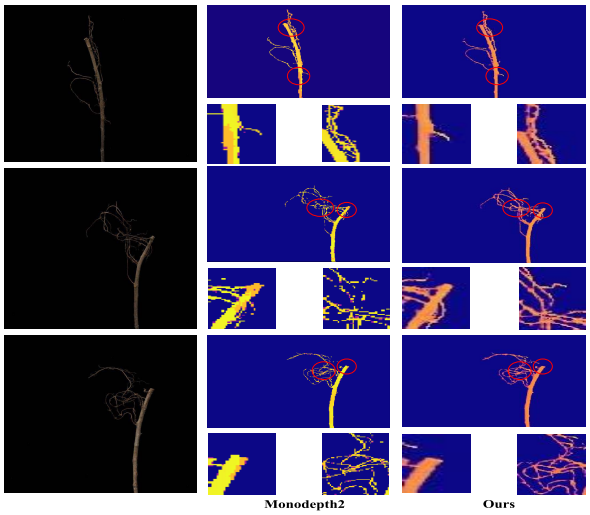

Fig. 13. Visual result of our method in disparity prediction from one single root image compared with [33]. Our method is able to generate more accurate and detailed prediction for thin and complicated root branch.

Fig. 14. Generalization performance of the proposed method to other objects (palm tree and wheat plant). The first column shows the original single image input. The second and third columns display the reconstructed plant models from different perspectives.

TABLE II

MEAN PSNR AND SSIM SCORES OF THE PROPOSED NOVEL VIEW SYNTHESIS METHOD COMPARED WITH RECENT GAN-BASED METHODS [30], [39], [45]–[47] ON THE COLLECTED DATASET

Method	PSNR	SSIM
Pix2Pix [45]	20.4	0.80
SingleGAN [39]	18.8	0.83
DualGAN [46]	23.6	0.85
Cross-net [30]	21.9	0.82
Pix2PixHD [47]	24.7	0.86
Ours (w/o multi-scale structure)	23.8	0.85
Ours (full pipeline)	25.9	0.87

TABLE III

MEAN DISTANCE AND VARIANCE TO THE GROUND TRUTH OF OUR METHOD COMPARED WITH OTHER SINGLE IMAGE DEPTH ESTIMATION AND MULTI-VIEW ESTIMATION METHODS ON THE COLLECTED DATASET. RESULTS DEMONSTRATE THAT THE MEAN DISTANCE AND VARIANCE OF OUR METHOD ARE SMALLER THAN [17], [26], [33], [43], [44]

Method	Mean Distance	Variance
Monodepth [26]	109.52	38.61
Monodepth2 [33]	37.69	12.47
Un-stereo-depthGAN [43]	66.78	15.76
MVDepthNet [17]	58.43	13.14
SC-SfMLearner [44]	59.63	14.10
Our direct depth estimation	35.92	12.65
Our synthetic stereo	33.26	9.79
Our full pipeline	29.42	9.23

TABLE IV

MEAN DISTANCE AND VARIANCE TO THE GROUND TRUTH OF OUR METHOD COMPARED WITH OTHER SINGLE IMAGE DEPTH ESTIMATION AND MULTI-VIEW ESTIMATION METHODS ON THE SYNTHETIC DATASET. RESULTS DEMONSTRATE THAT THE MEAN DISTANCE AND VARIANCE OF OUR METHOD ARE SMALLER COMPARED WITH [33], [44]

Method	Mean Distance	Variance
Monodepth2 [33]	33.47	14.16
SC-SfMLearner [44]	47.12	13.01
Our direct depth estimation	30.98	12.23
Our synthetic stereo	28.62	10.81
Our full pipeline	24.30	10.24

our method is able to produce the least distortion and stretching in shapes. Compared with and [44], our framework can preserve more details for thin branches.

The proposed method is also able to generalize to other normal objects in Figure 14. With the help of the proposed simultaneous direct depth estimation and synthesis stereo methods, the reconstructed point cloud from the single input image demonstrates a much more complete shape, especially on some thin branches. Also, it prevents some unexpected stretching and distortion usually found in each method. Beyond plant root systems, we further demonstrate that our pipeline is capable to reconstruct other common objects and scenes after fine-tuning the trained model on plant roots, as depicted in Fig. 15. We observe that although direct single image depth estimation can produce smoother regions on table surface and ground, synthetic stereo reconstruction can recover a complete model by persevering more details from another camera perspective. Results from our full pipeline take advantages of both depth estimation solutions.

2) Quantitative Comparisons: Numerical comparisons and analysis include PSNR and SSIM in the view prediction network with our method, and Mean distance with variance value (sigma) between point clouds for 3D reconstruction performance. As shown in Table II, our method achieves a good synthesis performance by improvement in PSNR from 4.9% to 37.7%, and SSIM from 1.2% to 8.8% over other recent methods on our real collected dataset, respectively.

To avoid the background affecting the evaluation, we compare foreground regions in the original images in Table II, III and IV quantitatively compares the average distance and variance of the generated point clouds from [17], [26], [33], [43], [44] and our method, with ground truth 3D models on the real collected dataset and Unity3D rendering on the synthetic dataset respectively. To evaluate the effectiveness of different methods better, we take the

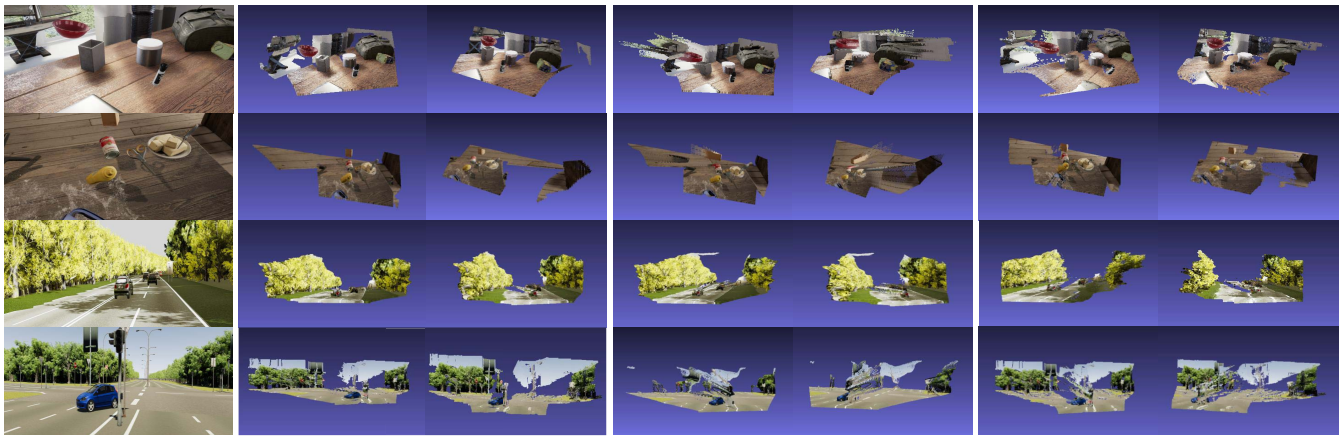

Fig. 15. Generalization on more common synthetic objects from [48] (the first two rows) and synthetic street scenes from [6] (the last two rows) beyond plant roots. Left to right: Input image; 3D reconstruction result from the full pipeline from two view perspectives; 3D reconstruction model from the direct single image depth estimation method only; 3D model from the synthetic stereo reconstruction method only.

mean distance as well as variance values together to measure the reconstruction quality. It can be found that the proposed method is capable of achieving smaller errors when compared to other methods.

V. Conclusion

In this paper, we have demonstrated a novel method to predict an accurate 3D point cloud from a single input image. We developed two different single image depth estimation methods, the direct depth estimation method and synthesis stereo reconstruction, which enjoy benefits from different perspectives. We fuse the 3D reconstruction models from these two methods by constraining their spatial consistency. Experiments demonstrate that the proposed pipeline significantly outperforms the state-of-the-art methods in reconstructing thin and complex plant roots from a single image. Furthermore, our method is truly self-supervised, which reduces the effort of labeling data. Although designed especially for plant roots, the proposed method is still able to be generalized to other common scenes like indoor objects and outdoor scenes.

References

[1] M. A. Khan, D. C. Gemenet, and A. Villordon, "Root system architecture and abiotic stress tolerance: Current knowledge in root and tuber crops," *Frontiers Plant Sci.*, vol. 7, p. 1584, Nov. 2016.

[2] E. D. Rogers and P. N. Benfey, "Regulation of plant root system architecture: Implications for crop advancement," *Current Opinion Biotechnol.*, vol. 32, pp. 93–98, Apr. 2015.

[3] A. X. Chang *et al.*, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*. [Online]. Available: http://arxiv.org/abs/1512.03012

[4] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, "Building generalizable agents with a realistic and rich 3D environment," 2018, *arXiv:1801.02209*. [Online]. Available: http://arxiv.org/abs/1801.02209

[5] Y. Wu, L. Jiang, and Y. Yang, "Revisiting EmbodiedQA: A simple baseline and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 3984–3992, 2020.

[6] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," 2020, *arXiv:2001.10773*. [Online]. Available: http://arxiv.org/abs/2001.10773

[7] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 605–613.

[8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 628–644.

[9] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3D object reconstruction from a single depth view with adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 679–688.

[10] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang, "Pixel2mesh: Generating 3D mesh models from single RGB images," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 52–67.

[11] Y. Lu, Y. Wang, and G. Lu, "Single image shape-from-silhouettes," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3604–3613.

[12] S. Fang, X. Yan, and H. Liao, "3D reconstruction and dynamic modeling of root architecture *in situ* and its application to crop phosphorus research," *Plant J.*, vol. 60, no. 6, pp. 1096–1108, Dec. 2009.

[13] Y. Zheng, S. Gu, H. Edelsbrunner, C. Tomasi, and P. Benfey, "Detailed reconstruction of 3D plant root shape," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2026–2033.

[14] C. N. Topp *et al.*, "3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 18, pp. E1695–E1704, Apr. 2013.

[15] Q. Shan, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Occluding contours for multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4002–4009.

[16] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 767–783.

[17] K. Wang and S. Shen, "MVDepthNet: Real-time multiview depth estimation neural network," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 248–257.

[18] S. Agarwal *et al.*, "Building rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[19] C. Sweeney, T. Sattler, T. Hollerer, M. Turk, and M. Pollefeys, "Optimizing the viewing graph for structure-from-motion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 801–809.

[20] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[21] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1983–1992.

[22] Y. Zou, Z. Luo, and J.-B. Huang, "Df-Net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 36–53.

[23] Y. Lu, Y. Wang, Z. Chen, A. Khan, C. Salvaggio, and G. Lu, "3D plant root system reconstruction based on fusion of deep structure-from-motion and IMU," *Multimedia Tools Appl.*, vol. 80, pp. 1–17, Jan. 2021.

[24] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[25] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Proc. CVPR*, Jun. 2011, pp. 3057–3064.
[26] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.
[27] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Single-view to multiview: Reconstructing unseen views with a convolutional network," 2015, *arXiv:1511.06702*. [Online]. Available: http://arxiv.org/abs/1511.06702
[28] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 286–301.
[29] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[30] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 867–875.
[31] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: http://arxiv.org/abs/1605.05396
[32] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.
[33] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[35] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[36] Y. Luo *et al.*, "Single view stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 155–163.
[37] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[38] K. Regmi and A. Borji, "Cross-view image synthesis using conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3501–3510.
[39] X. Yu, X. Cai, Z. Ying, T. Li, and G. Li, "Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 341–356.
[40] J. K. Haas, "A history of the unity game engine," MA, USA, Tech. Rep., 2014.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[43] A. Pilzer, S. Lathuilière, D. Xu, M. M. Puscas, E. Ricci, and N. Sebe, "Progressive fusion for unsupervised binocular depth estimation using cycled networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2380–2395, Oct. 2020.
[44] J.-W. Bian *et al.*, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1–11.
[45] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[46] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for Image-to-Image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.
[47] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[48] M. Jalal, J. Spjut, B. Boudaoud, and M. Betke, "SIDOD: A synthetic image dataset for 3D object pose recognition with distractors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 475–477.