2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 1-5, 2023. Detroit, USA

Bird-view 3D Reconstruction for Crops with Repeated Textures

Guoyu Lu

***Abstract*— Large-scale in-situ 3D reconstruction of crop fields presents a challenging task, as the 3D crop structures play a crucial role in plant phenotyping and significantly influence crop growth and yield. While existing efforts focus on close-range plants, only a limited number of deep learning-based methods have been developed explicitly for large-scale 3D crop reconstruction, mainly due to the scarcity of large-scale crop sensing data. In this paper, we leverage unmanned aerial vehicles (UAVs) in agriculture and utilize a recently captured multi-view real-world snap beans crop dataset to develop an unsupervised structure-from-motion (SfM) framework. Our framework is designed specifically for reconstructing large-scale 3D crop structures. It addresses the challenge of inaccurate depth inference caused by excessively repeated patterns in the crop dataset, resulting in highly accurate 3D crop reconstruction for large-scale scenarios. Through experiments conducted on the crop dataset, we demonstrate the accuracy and robustness of our 3D crop reconstruction algorithm. The application of our proposed framework has the potential to advance research in agriculture, enabling better plant phenotyping and understanding of crop growth and yield.**

I. INTRODUCTION

Large-scale highly accurate 3D structure of the farmland is important information regarding the holistic crop field structure, especially in precision agriculture. An accurate 3D reconstruction model of the crop provides special phenotypes for crop breeding and crop selection to be helpful with breeding next-generation crops [39][18]. 3D reconstructed crop models can be used to evaluate crop growth and crop yield, which allows further effective crop management. Therefore, the research for 3D reconstruction of the farmland is indispensable to greatly improve and speed up the ability of researchers and scientists to evaluate crop structural change, finally towards the objective of effective management of the large-scale crop fields. To obtain robust and accurate crop reconstruction, the current 3D crop reconstruction methods mainly use active sensing techniques (structured light [60][37], Time-of-flight (TOF) cameras [17], LiDAR [55][9]), and image-based 3D reconstruction techniques (Structure from Motion [41][30], multi-view stereo [55][4]) to acquire a single reconstructed plant. But research has rarely been conducted on large-scale farmland 3D modeling, which reduces the comprehensiveness of the overall crop analysis, leading to inefficient crop management. In addition, the above-mentioned sensing-techniques-based systems are very costly, which is not favorable to large-scale scene reconstruction applications. The above-mentioned image-based reconstruction methods are based on conventional methods, which generate inaccurate 3D reconstruction models due to inevitably mismatched feature points during reconstruction. In recent times, a limited number of deep learning approaches can produce satisfactory results on close-range plants [30][31][24], which target at ego-view individual plant reconstruction. However, they have a critical limitation when applied to large-scale farmland reconstruction using multiple-view crop images from a bird's eye view due to the repeated textures, which makes it difficult to train and improve an appropriate network for large-scale crop reconstruction. In addition, most of the existing crop datasets are captured in the laboratory environment, which is difficult to be directly applied to large-scale 3D crop reconstruction.

In this paper, we present an unsupervised 3D reconstruction framework that is applied to a real-world snap bean crop dataset captured by the UAV for large-scale 3D reconstruction. The dataset includes 15 snap bean crop sequences and camera intrinsics. The learning-based unsupervised SfM network focuses on exploiting multi-view geometric constraints to reconstruct large-scale and highly accurate 3D crop structures. The SfM network uses those constraints to estimate accurate UAV pose and depth maps. Due to the similar intensity and color of the crop field images, photometric loss may not effectively establish tight constraints. We extract regional contours in crop images to build relationships between successive images as the contours can distinctively represent the local crop shaping properties. To further overcome excessively repeated patterns in the crop dataset and enable the trained large-scale 3D farmland to maintain more details, we detect and match keypoints that constrain the consistency of the salient details across consecutive crop images as a loss to improve 3D crop structure modeling effects. The entire unsupervised SfM network explores spatial and spectral constraints. Spatial constraint is to enforce the contours and keypoints to be consistent across different frames on positions to reduce the displacement error to optimize the depth map, and spectral constraint is to enforce the pixel values to be consistent with pixel values corresponding to the same 3D point on different frames. With the support of both constraints, our proposed SfM network can achieve a superior large-scale 3D crop reconstruction effect.

Overall, the contributions of our work are summarized as follows: 1) we propose a learning-based unsupervised SfM network targeting at large-scale crop reconstruction to explore the crop properties; 2) we investigate the specific texture repetition properties of crops and propose regional contour consistency constraint to establish the relationship between consecutive frames; 3) the proposed method explores the multi-view geometric constraints to optimize and

Guoyu Lu is with the Intelligent Vision and Sensing (IVS) Lab at the University of Georgia, USA, guoyulu62@gmail.com

978-1-6654-9190-7/23/$31.00 ©2023 IEEE

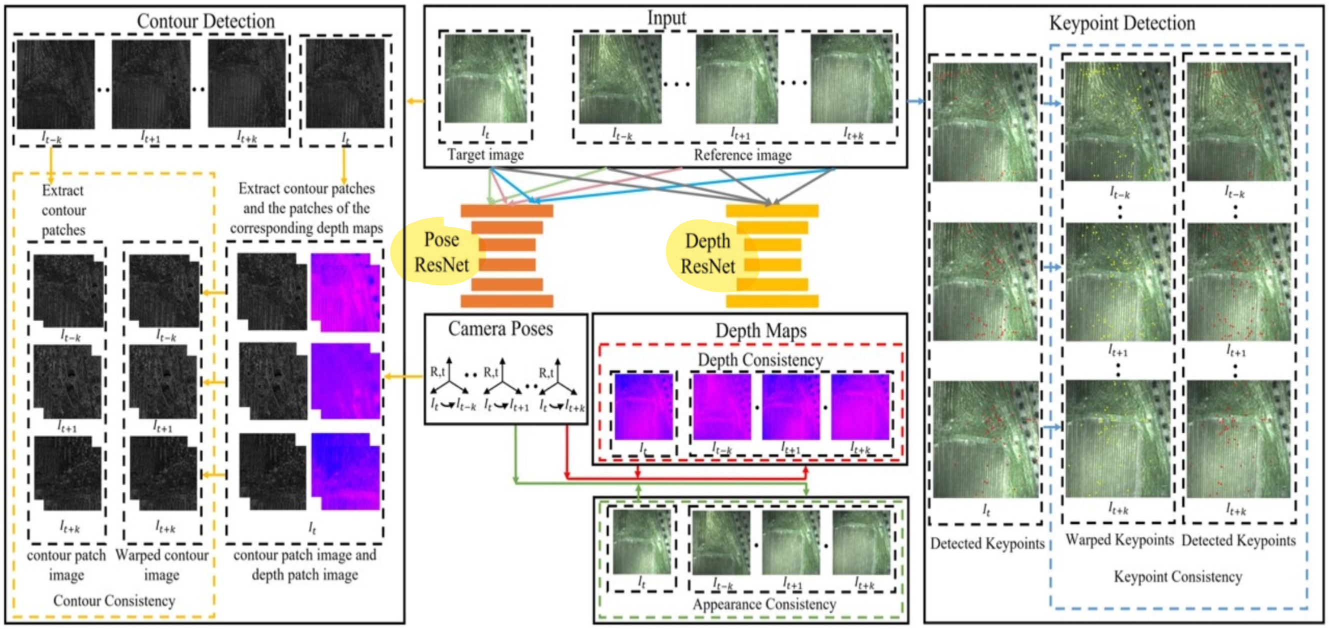

Fig. 1. The framework of our unsupervised 3D reconstruction method for large-scale crop field reconstruction, which explores the regional contour shape properties, keypoint matching relationship, and spectral consistency in the interaction between depth and pose estimation neural network branches.

enhance 3D crop structure effect. The entire reconstruction framework is shown in Fig. 1.

II. Related Work

Mainstream 3D crop reconstruction methods can be roughly divided into two categories: active sensing-based methods and passive image-based methods. Structured light (Kinect sensor), ToF cameras, and LiDAR are three major active sensing methods for 3D crop reconstruction. Structure-from-motion and multi-view stereo are two image-based methods for 3D reconstruction with only one or two cameras.

A. *Active Sensing-based Methods*

Structured light is a group of systems composed of a projector and a camera. After the projector projects specific light information on the surface of the object and the background, this light information is collected by the camera. According to the change of the light signal caused by the object, information such as the position and depth of the object can be calculated using triangulation [40] to restore the entire three-dimensional shape. Botterill et al. [3] applied a robot equipped with stereo cameras and a structured light scanner to acquire high-level 2D features and a sparse set of 3D structured light points. However, it did not offer the resolution to cope with complex plant structures. Li et al. [21] applied a structured light scanner with a camera-projector pair to obtain 4D reconstruction, with time being the fourth dimension. Nguyen et al. [37] developed a structured light system with the corresponding software algorithms to produce 3D models of entire plants without cutting away any parts of a plant destructively.

ToF camera uses active light to determine depth information. The sensor emits a light signal which hits the subject and returns to the sensor. The time it takes to bounce back is then measured, which provides depth-mapping capabilities. Kazmi et al. [17] evaluated the performance of ToF cameras for close-range plant images under three different illumination conditions and concluded that while ToF cameras can generate high frame rate accurate depth data under suitable conditions, the sensors are very sensitive to ambient light with low resolution, which causes the poor performance under outdoor environment. Therefore, farmland reconstruction is challenging under outdoor conditions using ToF cameras.

LiDAR, as a specific type of ToF camera, measures the distance between the object and the sensor using the laser, followed by calculating the time taken for the reflected light to come back [40]. Kaminuma et al. [15] presented precise 3D measurements using a laser range finder (LRF) and automatic data processing for phenotypic analysis of plants. Paulus et al. [42] proposed an approach automatically analyzing barley organs using 3D laser scanning to achieve the automated parameter tracking of the plant organs, leaf, and stem, but the point cloud data obtained by LiDAR loses the details of plants such as plant surface area. To further improve the accuracy of the model, Wu et al. [55] combined point cloud data generated by multiview sequence images as a reference to calibrate the point cloud data from the LiDAR scanning using Iterative Closest Point (ICP), finally establishing an accurate 3D model of the plant. However, LiDAR is very costly and needs calibration when used with cameras. In addition, LiDAR sensing is generally low resolution compared with normal cameras.

B. *Passive Image-based Methods*

Structure-from-Motion (SfM) is a widely used 3D reconstruction method for large-scale 3D modeling [52][49][45][28][29]. Classical depth prediction methods mainly rely on handcrafted features or probabilistic models either to estimate depth information from a single image [52][16][20][13] or from stereo images [57][33], which generally create a sparse 3D point cloud. Deep neural networks have been extensively used to estimate depth maps due to their superior ability to extract features from images [5][22][43]. Quan et al. [44] and Tan et al. [53] proposed a semi-automatic method for completed plant reconstruction using the SfM. Although those methods achieved good performance under outdoor conditions, it is difficult to reconstruct each leaf accurately. Jay et al. [13] applied the SfM method with RGB images acquired by a single camera moving along the rows of a crop field to reconstruct a crop 3D model.

Multi-view stereo (MVS) has a similar theory to SfM, but the MVS refines the SfM steps to generate the dense reconstruction. Therefore, most classic 3D reconstruction methods combined SfM and MVS to generate accurate and dense point clouds [26][48][59][2][58]. Schönberger et al. present a MVS system called COLMAP [50]. With the success of the learning-based stereo methods, some existing research attempted to apply CNNs to the MVS task. Ji et al. [14] proposed the end-to-end learning network designed for MVS, called SurfaceNet, by building colored voxel cubes outside the network to encode the camera parameters through perspective projection. However, SurfaceNet generates low-resolution objects and has a huge GPU memory consumption of 3D voxel. DeepMVS [12] generates a plane-sweep volume for each reference image, but this method is unrealistic for large-scale scenes due to the large memory consumption. Bundle adjustment and pose graph have been proven to be able to optimize neural networks and improve 3D reconstruction effects [27]. Liu et al. [23]proposed RED-Net, which adopts the idea of regularized 2D cost maps from [56] to effectively exploit neighborhood information, making RED-Net achieve large-scale and full-resolution reconstruction.

III. BIRD-VIEW 3D CROP RECONSTRUCTION

In this section, we elaborate on the learning-based contour matching guidance, which is trained to match contours between different frames to facilitate the 3D crop reconstruction tasks so that repeated pattern issues of crop images can be addressed during training. To further fine-tune the accuracy of crop structures, we investigate keypoint consistency constraints. The process of both contour matching guidance and keypoint consistency involves keypoint detection and descriptor learning. Once we obtain the contour matching and keypoint consistency, a designated end-to-end SfM network can be developed to navigate the large-scale 3D crop field reconstruction effectively. The other two consistency constraints (appearance consistency and depth consistency) are also fused into the crop field reconstruction framework to improve the reconstruction accuracy and enrich details of 3D crop structures.

A. Network Structure

The 3D crops reconstruction framework is to predict the depth maps and camera poses with consecutive frames as input. Fig. 1 demonstrates the basic structure of our deep unsupervised 3D SfM crop reconstruction network, which is to jointly learn the depth map and the corresponding camera pose by both contour consistency, keypoint consistency, and depth and appearance consistency constraints between the target image and reference image. The whole SfM framework consists of two deep neural networks: Pose ResNet and Depth ResNet. Depth ResNet in Fig. 1 is to generate a depth map with an encoder-decoder network structure. The encoder network extracts significant features from the input crop images, composed of seventeen convolutional layers and a single fully connected layer. The decoder network applies to skip connections [25] to further interpret those feature representations to generate a depth map. Pose ResNet has a similar network structure to Depth ResNet, but instead of generating depth maps, Pose ResNet outputs relative 6 DoF parameters, which can construct a rotation matrix $\mathrm{R}(3 \times 3)$ and a translation vector $\mathrm{t}(3 \times 1)$.

B. Contour Matching Learning Guidance

Our contour matching consists of three steps: contour extraction, keypoint detection, and description extraction. The contour extraction part is embedded into the contour matching network to extract the contours of the input crop images in real-time using the Sobel algorithm provided by Kornia [47]. Then a detector network, called RF-Det [51], takes contour images as input to detect the keypoints with a score map, an orientation map, and a scale map. The contour images will be cropped into multiple patches according to these three maps to be fed into the descriptor network to extract fixed-length feature vectors for matching.

1) Keypoint detection: We first construct multi-scale maps from input contour images to do the enhancement for low-texture images. The motivation is that the texture of contour images is relatively low compared to the RGB images, and the quality of the sharpness and appearance is also much lower than RGB images. Therefore, constructing multi-scale maps M can keep different level features to detect keypoints.

Inspired by LF-Net [38], since multi-scale maps M can represent pixel responses on multi-scales, high-response pixels will be chosen as keypoints, generating the keypoint score map M_{score} using two softmax operators. The orientation and scale maps can be calculated by applying convolutions on multi-scale maps M with two 1×1 kernels to separately generate multi-scale orientation maps M_{gra} and scale maps M_{scale}. Once these maps and keypoints are acquired, we can determine the matching precision.

2) Description extraction: Given keypoints, the patch descriptor can be trained to obtain descriptions for the target region. The descriptor network is similar to the Hard-Net [35], He et al. [10] and L2-Net [54]. The descriptor network includes seven convolution layers, and each convolution layer consists of a batch normalization and ReLU except for the last layer. During training, the patches generated by keypoint orientation and scale can affect the matching precision of the descriptor, so that patch consistency is treated as a constraint to improve the accuracy of the matching patches. During inference, the descriptor output is multiple matching 128×128 patches P between two different frames, which can overcome the repeated pattern issues caused by crop images

to improve the accuracy of 3D crop reconstruction.

In the process of training for 3D crops reconstruction, as shown in the left part of Fig. 1, multiple matched patches between target image I_t and reference images $I_{t-k} \sim I_{t+k}$ are found, and the pixel coordinates corresponding to each patch can be extracted in original contour image. Meanwhile, the depth values corresponding to each patch will also be extracted. So we further enforce the coordinate points of multiple contour patches cropped from target contour images I_t to be consistent with coordinate points of reference contour images $I_{t-k} \sim I_{t+k}$ based on the depth values corresponding to each patch and camera motion. The proposed contour matching consistency constraints can be formulated as:

$$L_{\text{contour}} = \sum_M \sum_N \sum_i \left\| pat_{M,i}\left(\pi\left(P_{tr}, D_M\right)\right) - pat_{N,i} \right\|_2 \tag{1}$$

where M represents the number of target images. N represents the number of all other reference images except for the current target view. i is the number of extracted coordinate points from the contour patches. π represents a mapping relationship from the coordinate points from target contour patches to the coordinate points from other matched reference contour patches. P_{tr} is the relative camera motion from the target image to the reference images. D_M corresponds to the depth values at the current M_{th} target contour region. pat means the pixel coordinate points of contour patches. The $L2$ norm is to measure the distance between warped contour coordinate points in target contour patches and contour coordinate points in the reference images and minimize it.

C. Keypoint Consistency

The basic principle of the keypoint consistency constraint is to find the matching keypoints between target views and reference views to fine-tune the 3D crop structure. Therefore, we still follow the matching process of section III-B to learn matching keypoints, but different from contour matching, RGB crop images are taken as input to find matching keypoints in similar texture RGB images. Once we obtain the pre-trained model for matching keypoints, the model can be applied to 3D crop reconstruction. The right part of Fig. 1 presents the keypoint consistency constraint. With any consecutive RGB images as input, the detected matching keypoints between any two frames with the predicted depth information for each frame and the estimated relative camera motions between them can build an unsupervised constraint to enforce the warped keypoints of the target image to be consistent with detected keypoints of reference images. Therefore, the proposed keypoint consistency constraint is defined as:

$$L_{\text{keypoint}} = \sum_M \sum_N \sum_i \left\| poi_{M,i}\left(\pi\left(P_{tr}, D_M\right)\right) - p_{oi_{N,i}} \right\|_2 \tag{2}$$

where M, N, i, π, P_{tr} and D_M represent the same meaning as in Eq. 1. p_{oi} means the coordinate points of the detected keypoints. Similar to the contour matching guidance scheme, the $L2$ norm is also to minimize the difference between warped keypoints in the target image and the detected keypoints in the reference images.

D. Appearance Consistency

Based on contour consistency and keypoint consistency, we further add spectral consistency constraints to improve the details of 3D crop reconstruction, which enforces the spectral appearance of the warped target images to be consistent with reference images according to predicted crop depth maps and estimated camera poses. The spectral consistency constraints can be achieved as:

$$L_{spectral} = \sum_M \sum_N \left\| I_M\left(\pi\left(P_{tr}, D_M\right)\right) - I_N \right\|_1 \tag{3}$$

where M and N respectively represent the number of the target image and reference images, but the reference image is not the same as the current target image in each loss calculation. P_{tr} is the relative camera motion from the target image to the reference images. D_M corresponds to the depth map at current M_{th} target image. I_N is the N_{th} reference image. The L1 loss is applied to reduce the pixel RGB value difference between the warped target image and reference images.

As the robustness of L1 loss is not enough for the light illumination and contrast variation, the image structural similarity index (SSIM) is fused into the L_{spectral} to evaluate the similarity between two images in illuminance, contrast, and structure. The improved spectral consistency loss can be expressed by a combination of SSIM and L1 loss as:

$$\begin{aligned} L_{\text{spectral}} = & \sum_M \sum_N \lambda_1 \left\| I_M\left(\pi\left(P_{tr}, D_M\right)\right) - I_N \right\|_1 \\ & + \lambda_2 \frac{1 - \text{SSIM}\left(I_M, I_N\right)}{2} \end{aligned} \tag{4}$$

where $\text{SSIM}\left(I_M, I_N\right)$ computes the element-wise similarity between the warped target image I_M and the reference image I_N. We set $\lambda_1 = 0.15$ and $\lambda_2 = 0.85$ following [7] [6].

E. Depth Consistency

Depth represents and consists of the geometric information of images. Depth maps are less sensitive to the gradient locality [1] compared with color images. Therefore, we further introduce depth consistency across consecutive frames to solve possible depth ambiguity. Target depth map D_M can be warped into reference depth map $\tilde{D_N}$, called warped reference depth map, and then we force warped reference depth map to be consistent with the original reference depth map. A scaling ratio of both two depth maps is first calculated, and then the depth consistency constraint is defined as follows:

$$\begin{aligned} L_{\text{depth}} &= \sum_i \left| \eta \cdot \tilde{D}_N(i) - D_N(i) \right|, \\ \eta &= \frac{\sum_i D_N(i)}{\sum_i \tilde{D}_N(i)} \end{aligned} \tag{5}$$

where η is the depth scale ratio between the warped reference depth map and the original reference depth map. The integrated constraints of our framework are as follows:

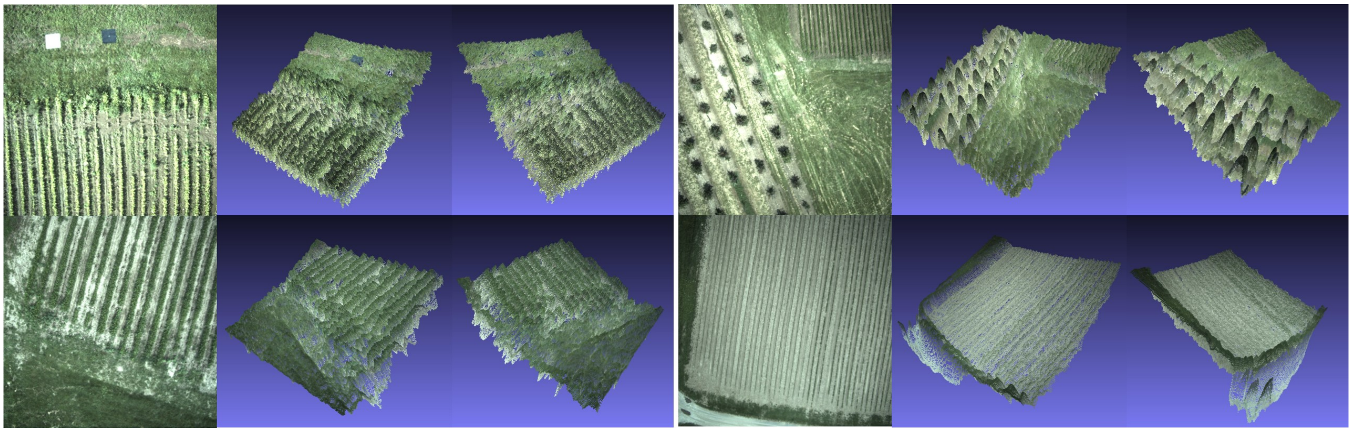

Fig. 2. 3D models from different perspectives reconstructed by our method based on a single image input.

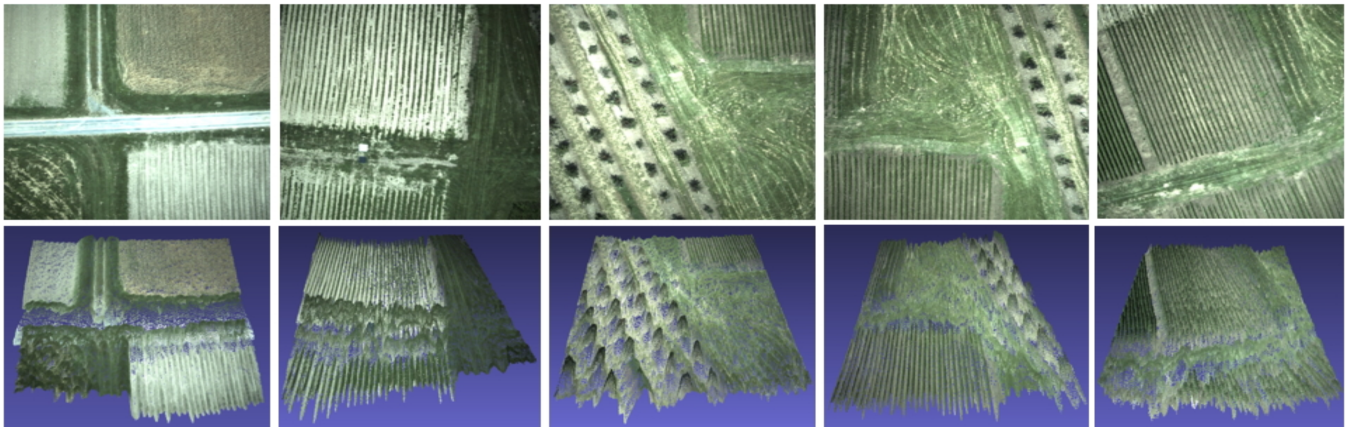

Fig. 3. 3D reconstruction models (below) based on a single image input (above).

$$L_{\text{total}} = \lambda_1 L_{\text{contour}} + \lambda_2 L_{\text{keypoint}} + \lambda_3 L_{\text{spectral}} + \lambda_4 L_{\text{depth}} \quad (6)$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are different weights for the four constraints.

IV. EXPERIMENT RESULTS AND ANALYSIS

A. Dataset

To generate large-scale 3D crop reconstruction, we use the Mako G-419 camera mounted on UAV to take consecutive crop images with recorded real-time LiDAR data. Although the Mako G-419 camera outputs multi-spectral images, we only processed the RGB channels using Spectral Python (SPy) toolbox with the resolution of 512×512. The calibration from LiDAR to Mako G-419 camera can be completed offline to obtain the calibration matrix. Therefore, the 3D data captured by LiDAR can be treated as ground truth to validate our model. We collected 15 snap bean sequences with 18,000 images.

B. Training Configuration

For contour matching and keypoint matching training, the input images are respectively contoured images through the Sobel operator and RGB images with the same size of 512×512. We first put these two types of images into the detector network called RF-Det to extract the high-response pixels as keypoints with three maps. These patches are determined by three maps input descriptor networks to extract fix-length feature vectors for matching. Although these two tasks have the same training process, their training configuration is different. For contour matching, in order to generate larger patches with fewer keypoints to extract contour patches, we set the patch size as $128 * 128$ and the top-k contour patch number as 32. As a comparison, we also want to generate more matched keypoints with smaller patches to detect more matched keypoints, for which we set the patch size as $32 * 32$ and the top-k number to 628.

After detecting matched contour patches and matched keypoints between the target image and reference images, the learning-based 3D crop reconstruction network is trained in an unsupervised manner and does not need any 3D supervision to guide the training process. It is implemented with PyTorch library and trained from scratch using Adam optimizer [19] with $\beta 1 = 0.9$ and $\beta 2 = 0.99$. Rectified Linear Units (ReLu) [36] is applied as activation functions for all convolutional layers. The weights of the depth estimation network and pose estimation network are initialized with the Kaiming initialization [11] by setting the batch size to 4 to achieve a trade-off between the efficiency and the memory usage. The whole framework is trained for 30 epochs.

C. 3D Visual Reconstruction Result

We first show the 3D modeling result based on a single image input, as Fig. 2. One can notice that the 3D modeling results are accurate from different perspectives with just an image input. Even with different crop types in the same field, the reconstruction models of the different crops are

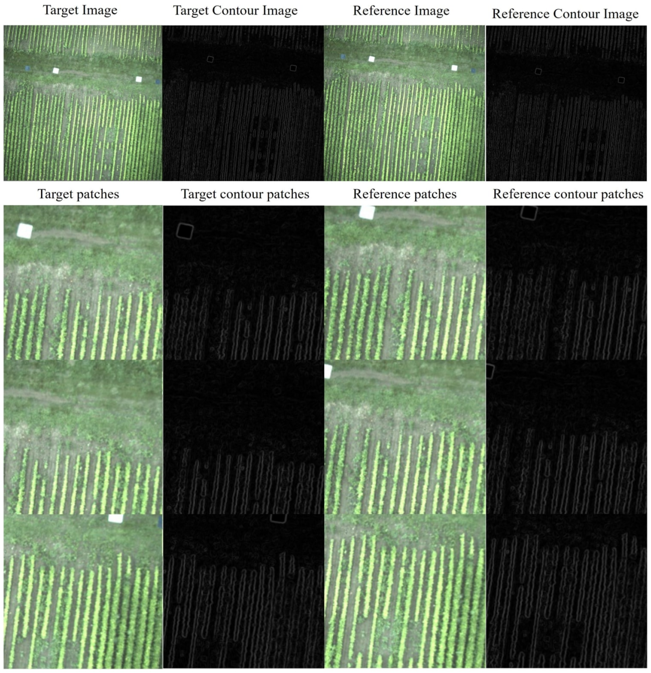

Fig. 4. First row: from left to right are the input target image, the corresponding target contour image, the reference image and the corresponding reference contour image; Last three row: the image patch and contour patch of the first two columns match the image patch and contour patch of the latter two columns respectively.

Metric	Contour Matching	Keypoint Matching
NN	0.400	0.329
NNT	0.511	0.482

TABLE I

Nearest neighbor (NN) and nearest neighbor with a threshold (NNT) on evaluating contour matching and keypoint matching methods.

still clearly distinguishable. More results are shown in Fig. 3, which also has shown accurate 3D crop reconstructions.

D. Contour Matching

Through training a set of consecutive crop contour images, the contour matching model has superior performance with matched patches. We visually present matching contour patches between the target image and the reference image. As shown in Fig. 4, all the corresponding contour patches are matched correctly, which demonstrates the target contour patches can match reference contour patches well during the training process to address the repeated pattern issues. Meanwhile, to evaluate the performance of our contour matching strategy, we use two matching metrics according to [34] to quantitatively calculate the matching score. The first one is nearest neighbor (NN) based matching. Supposing the descriptors of the target contour patches are the nearest neighbor to the descriptors of the reference contour patches, two contour patches match. Each descriptor has only a unique match. The second one is the nearest neighbor with a threshold (NNT) based matching. If the descriptors of the target contour patches are the nearest neighbor to the descriptors of the reference contour patches and the distance between two patches is less than a threshold, the two contour patches match. Correct matching ratios under these matching criteria are reported. As the first column of Table I, our quantitative results are respectively 0.4 and 0.511 on NN and NNT metrics.

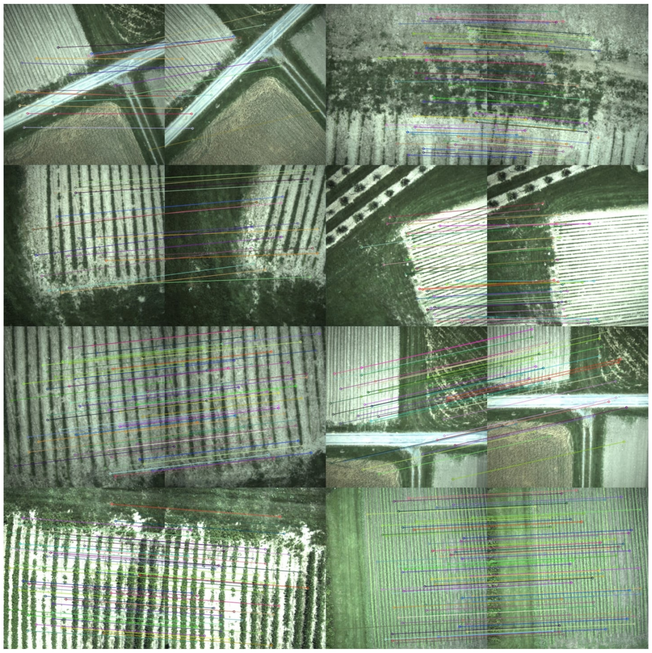

Fig. 5. Aligned 2D keypoints between the target and reference images

E. Keypoint Matching

Similarly, through training a set of consecutive crop images, the keypoint matching model represents superior performance with matched keypoints. We visually plot corresponding keypoints detected from the keypoint detector between the target image and reference images. As shown in Fig. 5, all the corresponding keypoints between two frames are matched correctly without being affected by any transformation such as rotation, translation, affine transformation and so forth. The results of the first two columns of the first row and the last two columns of the second row can show that our keypoint matching model can match correctly under any transformation. Meanwhile, we also use the same metrics as the contour matching method to validate our keypoint matching performance. As shown in the second column of Table I, our quantitative results are respectively 0.329 and 0.482 on NN and NNT metrics.

F. Comparison with other methods

As shown in Fig. 6, we compare our method with other state-of-the-art methods including PackNet-SfM [8], Monodepth2 [7], Vision Transformer [46], and HR-Depth [32]. The first column is the raw input images, and the second column is the 3D maps generated by our method, which reflects the depth information of the crop image, and the map is not distorted. But for the other models, the output has severe distortions. PackNet-SfM can capture some texture information, but the 3D depth maps are still obviously distorted. The remaining methods can no longer output meaningful 3D crop field models.

In the quantitative analysis, we use the 3D points data collected by LiDAR as the ground truth and calculate the error

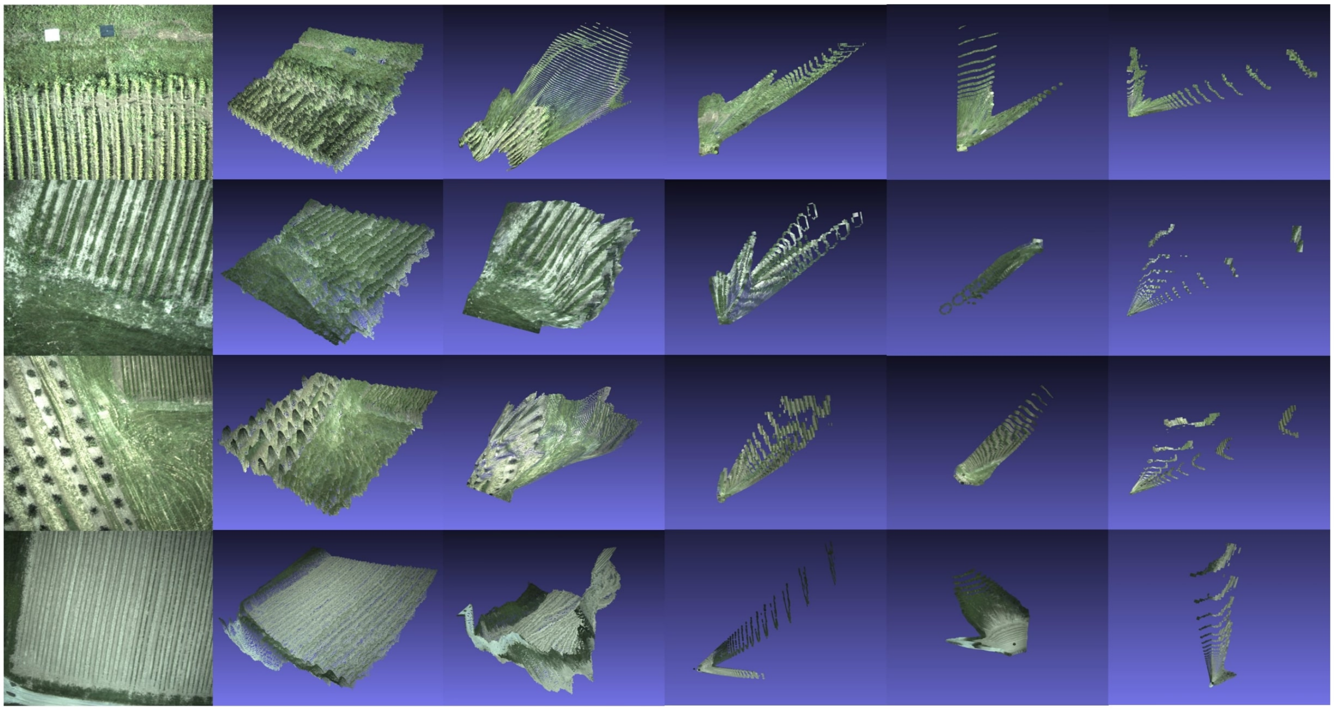

Fig. 6. Visual comparisons on reconstructed crop shapes from snap bean crop datasets between our result and other recent 3D reconstruction methods. First column: raw input image; second column: result from our pipeline; third column: result from PackNet-SfM method [8]; fourth column: result from Monodepth2 [7] method; fifth column: result from Vision Transformer [46] method; sixth column: result from HR-Depth [32] method.

Metric	Our Method	HR-Depth	Monodepth 2	PackNet-SfM	Vision Transformer
MAE	0.418	0.861	0.972	0.812	0.728
MSE	0.376	0.777	0.851	0.682	0.594

TABLE II

Mean Average Error (MAE) and Mean Squared Error (MSE) on snap bean dataset.

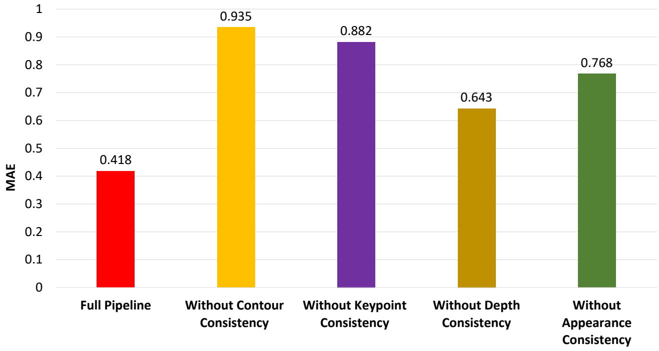

Fig. 7. Ablation study for the full pipeline and network with specific consistency constraints. The measurement is based on MAE.

of the depth map generated by different models compared to the ground truth. The results are shown in Table II. In both MAE and MSE metrics, our method is significantly better than the others.

We also conducted an ablation study to compare the full pipeline with the pipeline missing each specific consistency constraint, as shown in Fig. 7. One can see that each component contributes to the depth estimation output. Among all those consistency constraints, contour consistency contributes the most to accurate 3D crop modeling, followed by keypoint consistency, appearance consistency, and then depth consistency. This has demonstrated that the network is effective for 3D crop modeling tasks.

V. Conclusion

This paper proposes an unsupervised SfM neural network designated for 3D crop field reconstruction, which is typically difficult for both conventional and learning-based SfM methods due to the repeated textures and similar colors and intensities across images. To resolve this issue, we investigate contour-based and keypoint matching consistency constraints to guide the learning process together with the appearance and depth consistencies. Based on both spatial and spectral constraints, the proposed neural network can effectively reconstruct large-scale crop fields based on bird-view images captured by UAVs.

Acknowledgement

This publication is based upon work supported by NSF under Awards No. 2104032 and 2105257, and USDA NIFA grant No. 2021-67021-34199.

References

[1] P Anandan, James R Bergen, Keith J Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. *Motion analysis and image sequence processing*, pages 1–22, 1993.

[2] Dionisio Andújar, Mikel Calle, César Fernández-Quintanilla, Ángela Ribeiro, and José Dorado. Three-dimensional modeling of weed plants using low-cost photogrammetry. *Sensors*, 18(4), 2018.

[3] Tom Botterill, Richard Green, and Steven Mills. Reconstructing partially visible models using stereo vision, structured light, and the g2o framework. In *ISVCZ*, pages 370–375, 2012.

[4] Daniel Cremers and Kalin Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE PAMI*, 33(6):1161–1174, 2010.

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015.

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017.

[7] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

[8] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020.

[9] Qinghua Guo, Fangfang Wu, Shuxin Pang, Xiaoqian Zhao, Linhai Chen, Jin Liu, Baolin Xue, Guangcai Xu, Le Li, Haichun Jing, et al. Crop 3d—a lidar based platform for 3d high-throughput crop phenotyping. *Science China Life Sciences*, 61:328–339, 2018.

[10] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *CVPR*, pages 596–605, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015.

[12] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, pages 2821–2830, 2018.

[13] Sylvain Jay, Gilles Rabatel, Xavier Hadoux, Daniel Moura, and Nathalie Gorretta. In-field crop row phenotyping from 3d modeling performed using structure from motion. *CEA*, 110:70–77, 2015.

[14] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *ICCV*, pages 2307–2315, 2017.

[15] Eli Kaminuma, Naohiko Heida, Yuko Tsumoto, Naoki Yamamoto, Nobuharu Goto, Naoki Okamoto, Akihiko Konagaya, Minami Matsui, and Tetsuro Toyoda. Automatic quantification of morphological traits via three-dimensional measurement of arabidopsis. *The Plant Journal*, 38(2):358–365, 2004.

[16] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth extraction from video using non-parametric sampling. *arXiv preprint arXiv:2002.04479*, 2019.

[17] Wajahat Kazmi, Sergi Foix, Guillem Alenyà, and Hans Jørgen Andersen. Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: Analysis and comparison. *P&RS*, 88:128–146, 2014.

[18] Young Cheol Kim, Johan Leveau, Brian B McSpadden Gardener, Elizabeth A Pierson, Leland S Pierson III, and Choong-Min Ryu. The multifactorial basis for plant health promotion by plant-associated bacteria. *AEM*, 77(5):1548–1555, 2011.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Janusz Konrad, Meng Wang, and Prakash Ishwar. 2d-to-3d image conversion by learning depth from examples. In *CVPRW*, 2012.

[21] Yangyan Li, Xiaochen Fan, Niloy J Mitra, Daniel Chamovitz, Daniel Cohen-Or, and Baoquan Chen. Analyzing growing plants from 4d point cloud data. *ACM TOG*, 32(6):1–10, 2013.

[22] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015.

[23] Jin Liu and Shunping Ji. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In *CVPR*, pages 6050–6059, 2020.

[24] Zhihao Liu, Kai Wu, Jianwei Guo, Yunhai Wang, Oliver Deussen, and Zhanglin Cheng. Single image tree reconstruction via adversarial network. *Graphical Models*, 117:101115, 2021.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[26] Lu Lou, Yonghuai Liu, Jiwan Han, and John H Doonan. Accurate multi-view stereo 3d reconstruction for cost-effective plant phenotyping. In *ICIAP*, pages 349–356, 2014.

[27] Guoyu Lu. Deep unsupervised visual odometry via bundle adjusted pose graph optimization. In *ICRA*, pages 6131–6137, 2023.

[28] Guoyu Lu, Vincent Ly, and Chandra Kambhamettu. Large-scale structure-from-motion reconstruction with small memory consumption. In *MoMM*, pages 500–508, 2013.

[29] Guoyu Lu, Vincent Ly, and Chandra Kambhamettu. Structure-from-motion reconstruction based on weighted hamming descriptors. In *IJCNN*, pages 2367–2374, 2014.

[30] Yawen Lu, Yuxing Wang, Zhanjie Chen, Awais Khan, Carl Salvaggio, and Guoyu Lu. 3D plant root system reconstruction based on fusion of deep structure-from-motion and IMU. *MTA*, 80:17315–17331, 2021.

[31] Yawen Lu, Yuxing Wang, Devarth Parikh, Awais Khan, and Guoyu Lu. Simultaneous direct depth estimation and synthesis stereo for single image plant root reconstruction. *IEEE TIP*, 30:4883–4893, 2021.

[32] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *AAAI*, 2021.

[33] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016.

[34] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27(10):1615–1630, 2005.

[35] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *NIPS*, 30, 2017.

[36] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010.

[37] Thuy Tuong Nguyen, David C Slaughter, Nelson Max, Julin N Maloof, and Neelima Sinha. Structured light-based 3d reconstruction system for plants. *Sensors*, 15(8):18587–18612, 2015.

[38] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *NIPS*, 31, 2018.

[39] Martin AJ Parry, Pippa J Madgwick, Carlos Bayon, Katie Tearall, Antonio Hernandez-Lopez, Marcela Baudo, Mariann Rakszegi, Walid Hamada, Adnan Al-Yassin, Hassan Ouabbou, et al. Mutation discovery for crop improvement. *JXB*, 60(10):2817–2825, 2009.

[40] Abhipray Paturkar, Gourab Sen Gupta, and Donald Bailey. Overview of image-based 3d vision systems for agricultural applications. In *IVCNZ*, pages 1–6, 2017.

[41] Abhipray Paturkar, Gaurab Sen Gupta, and Donald Bailey. 3d reconstruction of plants under outdoor conditions using image-based computer vision. In *RTIP2R*, pages 284–297, 2019.

[42] Stefan Paulus, Jan Dupuis, Sebastian Riedel, and Heiner Kuhlmann. Automated analysis of barley organs using 3d laser scanning: An approach for high throughput phenotyping. *Sensors*, 14(7):12670–12686, 2014.

[43] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, pages 283–291, 2018.

[44] Long Quan, Ping Tan, Gang Zeng, Lu Yuan, Jingdong Wang, and Sing Bing Kang. Image-based plant modeling. In *ACM Siggraph*, pages 599–604. 2006.

[45] Changchang Wu Raguram, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building rome on a cloudless day. In *ECCV*, 2010.

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

[47] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *WACV*, pages 3674–3683, 2020.

[48] Johann Christian Rose, Stefan Paulus, and Heiner Kuhlmann. Accuracy analysis of a multi-view stereo approach for phenotyping of tomato plants at the organ level. *Sensors*, 15(5):9651–9665, 2015.

[49] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE TPAMI*, 31(5):824–840, 2008.

[50] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518, 2016.

[51] Xuelun Shen, Cheng Wang, Xin Li, Zenglei Yu, Jonathan Li, Chenglu Wen, Ming Cheng, and Zijian He. Rf-net: An end-to-end image matching network based on receptive field. In *CVPR*, 2019.

[52] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM Siggraph*. 2006.

[53] Ping Tan, Gang Zeng, Jingdong Wang, Sing Bing Kang, and Long Quan. Image-based tree modeling. In *ACM SIGGRAPH*. 2007.

[54] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *CVPR*, 2017.

[55] Jingwen Wu, Xinyu Xue, Songchao Zhang, Weicai Qin, Chen Chen, and Tao Sun. Plant 3d reconstruction based on lidar and multi-view sequence images. *IJPAA*, 1(1), 2018.

[56] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, pages 5525–5534, 2019.

[57] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*, 17(1):2287–2318, 2016.

[58] Kai Zhang, Noah Snavely, and Jin Sun. Leveraging vision reconstruction pipelines for satellite imagery. In *ICCVW*, 2019.

[59] Yu Zhang, Poching Teng, Yo Shimizu, Fumiki Hosoi, and Kenji Omasa. Estimating 3d leaf and stem shape of nursery paprika plants by a novel multi-camera photography system. *Sensors*, 16(6), 2016.

[60] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.