Multimedia Tools and Applications (2021) 80:17315–17331
https://doi.org/10.1007/s11042-020-10069-3

3D plant root system reconstruction based on fusion of deep structure-from-motion and IMU

Yawen Lu[1] · Yuxing Wang[1] · Zhanjie Chen[1] · Awais Khan[2] · Carl Salvaggio[1] · Guoyu Lu[1]

Received: 3 May 2020 / Revised: 6 September 2020 / Accepted: 7 October 2020 /
Published online: 2 January 2021

Abstract

Roots play a critical role in the functioning of plants. However, it is still challenging to generate detailed 3D models of thin and complicated plant roots, due to the complexity of the structure and the limited textures. Limited by the difficulty of realization and inaccessibility of labeled data for training, few works have been put in exploring this problem using deep neural networks. To overcome this limitation, this paper presents a structure-from-motion based deep neural network structure for plant root reconstruction in a self-supervised manner, which can be applied by mobile phone platforms. In the training process of deep structure-from-motion, each depth is constrained from the depth map and predicted relative poses from their adjacent frames captured by the mobile phone cameras, and the LSTM-based network after CNN for pose estimation is learnt from the ego-motion constraints by further exploiting the temporal relationship between consecutive frames. IMU unit in the mobile phone is further utilized to improve the pose estimation network by continuously updating the correct scales from the gyroscope and accelerometer moment. Our proposed approach is able to solve the scale ambiguity in recovering the absolute scale of the real plant roots so that the approach can promote the performance of camera pose estimation and scene reconstruction jointly. The experimental results on both real plant root dataset and the rendered synthetic root dataset demonstrate the superior performance of our method compared with the classical and state-of-the-art learning-based structure-from-motion methods.

Keywords Structure-from-motion · Convolutional neural network · IMU · Fusion

✉ Guoyu Lu
luguoyu@cis.rit.edu

Awais Khan
awais.khan@cornell.edu

[1] Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, USA

[2] Plant Pathology and Plant-Microbe Biology Section, Cornell University, Ithaca, NY, USA

1 Introduction

Roots are the main conduit for plants interaction with its physical and biological environment. Roots play a key role in crops and plants to draw nutrients from the soil to survive and develop [29, 47], and reconstructed 3D root models are able to provide fundamental evidences for growers and biologists to measure various traits including root length, volumetric biomass, and crop growth [8, 9]. However, the plant root system is very complicated, with a lot of thin and small branches with similar appearance and colors, which is very difficult to reliably extract unique features and reconstruct them in the 3D model. Recent approaches focusing on generating 3D shapes from a single image or multiple images [10, 13, 58, 66] benefit from deep learning techniques to enhance the reconstruction accuracy. The inferred output from these approaches could be expressed as the form of point clouds [13, 66] or a volume [10, 58]. However, such methods mentioned above heavily rely on the geometric representation from ground truth 3D models or CAD information, which is not feasible to scan each object in real applications. Also, though existing methods are able to deal with normal objects like cars or chairs, they fail to reconstruct plenty of small and detailed breaches in plant roots.

Multi-modal fusion technique as an import task in computer vision and robotics [53], which has been explored in a variety of tasks such as path planning, robotics controlling and video understanding [56, 62, 68]. Leveraging the benefits of each modality sensors, multi-modal fusion would achieve more robustness and better performance in the entire framework than applying any one of them alone as each sensor maintains its own limitations. Previous attempts in integrating visual cameras and inertial sensors [7, 26, 31, 32] enable the mobile devices to provide more accurate pose estimation. Such two sensors are relatively cheap and portable, thus are widely equipped in mobile devices like mobile phones. Single Camera is capable to extract pixel-level information from images and use this information for matching and 3D reconstruction. However, it fails to recover the absolute scale in the 3D world. Also they are not robust for a variety of illumination changes or high motion scenarios. IMU as a useful device for navigation, it avoids being affected by the scale ambiguity, but still suffers from accumulated errors and bias. Current approaches like [7, 31] mainly rely on hand-crafted features and aim to fuse them together either based on Extended Kalman Filter (EKF) [7] or full smoothers [31]. Different from these works, we explore to apply a deep neural network to leverage their benefits jointly.

In this paper, we introduce a self-supervised deep structure-from-motion (SfM) based system with inertial measurement to reconstruct a complete 3D plant root system structure from multiple images as input during the real application, as depicted in Fig. 1. We introduce a novel sensor fusion strategy to incorporate synchronized motion and rotation information from the raw inertial data of mobile devices. The combination of visual camera and IMU sensors is able to correct the scale issue of the camera motion estimation in monocular SfM in real-time. Furthermore, we extend the individual frame's depth map from the learning-based SfM network to a complete point cloud using a depth fusion strategy to recover a completed 3D structure of the root system. Additionally, we contribute both a real IMU-visual plant root dataset and a virtually rendered dataset to verify the performance and robustness of our framework and provide a comparison between our framework and state-of-the-art methods. In summary, the main contributions of this work are listed as follows: 1. We develop a novel deep structure-from-motion framework to recover a complete 3D point cloud in a self-supervised manner. 2. We propose a sensor fusion strategy to correct the scale ambiguity for monocular SfM by the camera and the IMU. 3. We build both real and

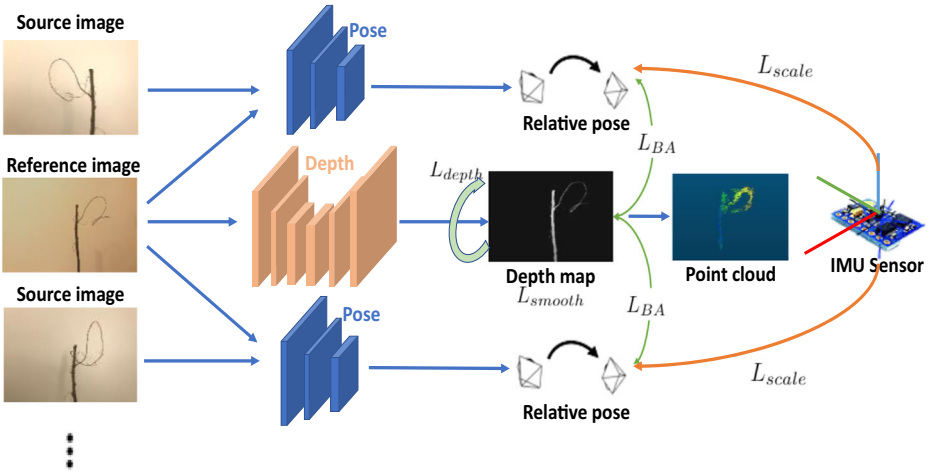

Fig. 1 Overview of overall training framework. Our proposed network takes image sequences as input for training, and simultaneously refine the intermediate camera motion as well as the depth maps to output a complete point cloud. IMU measurement is supplied and utilized as an additional input to dynamically refine the scale from the pose estimation network to provide more accurate estimation

synthetic plant root datasets to verify the 3D reconstruction performance of our framework, compared with other recent learning-based approaches and traditional SfM methods.

2 Related work

2.1 Classical SfM methods

Classical structure from motion (SfM) methods refer to a process of 3D reconstruction and camera pose estimation from multiple ordered or unordered images from different perspectives [1, 2, 24, 42, 51]. Those conventional SfM approaches recover the 3D scenes by incrementally and continuously adding new images to the scene with a process of feature extraction, feature matching and geometric verification together, followed by refining the results of reconstruction using RANSAC [17] and bundle adjustment methods [54, 64] which optimize the global or local structure. The early systems applied to the small scale 3D reconstruction and camera pose estimation were implemented based on the uncalibrated or self-calibrated metric reconstruction system [4, 12, 15, 44]. Snavely et al. [50] realized city-scale 3D SfM reconstruction based on images downloaded from Internet. However, when more and more cameras are added, this SfM method will be significantly slower. In addition, the repeated scene structure will produce many wrong matches. Therefore, more large city-scale SfM reconstruction methods on the basis of [50] have also been developed using the novel distributed algorithms for feature matching and 3D points' position estimation, for saving the computational cost [2, 18, 37, 45, 48, 49, 63, 69]. Especially, Changchang Wu improved the time complexity of the SfM algorithm to $o(n^2)$ compared with $o(n^4)$ of [50] by using a preemptive feature matching strategy and retained high accuracy at the same time. Other SfM strategies have been proposed including global, hybrid, hierarchical SfM reconstruction methods. Compared with incremental models producing drift due to cumulative errors, global reconstruction models do not drift as significant as incremental methods

because of the optimization of the spatial relationship between cameras [27, 38, 61, 73]. However, global optimization usually takes quite long time to accurately reconstruct all the images. Hybrid reconstruction models combine the advantages of incremental and global SfM to a certain extent to balance the efficiency and accuracy of reconstructed scenes [11, 72]. Hierarchical reconstruction models attempt to solve the problems of initialization and drift from incremental models based on hierarchically add new images to reconstruct [14, 20]. However, the existing classical approaches depend on the accurate feature extraction and matching, which is not always ideal in natural scenes, such as occlusions, few textures, low illumination, dark and foggy scenes.

2.2 Learning-based reconstruction from images

Recent reconstruction systems applying deep neural networks depend heavily on ground truth labels to generate 3D models from a single image or multiples images [10, 13, 33, 35, 36, 55, 58, 66]. Choy et al. [10] proposed a 3D Recurrent Reconstruction Neural Network (3D-R2N2) to input one image or multiple viewpoint images of an object and train a mapping from images to their 3D shapes. Fan et al. [13] took a single arbitrary depth view as input to reconstruct a full complete 3D occupancy grid of an object using generative adversarial networks [66]. Ummenhofer et al. [55] train a convolutional network to formulate structure from motion as a learning problem from successive and unconstrained image pairs. However, the most significant limitation for the aforementioned data-driven and offline trained systems is that the ground truth labels are not always available and feasible in real-world scenarios.

Unsupervised learning networks are also proposed to estimate depth map and/or camera odometry based on intensity consistency between input image pairs [3, 16, 19, 21, 41, 43, 59, 67, 70]. Among them, the network developed by Zhou et al. [71] firstly learned single-view depth and ego-motion with a loss calculation of the warp images of the views around the target jointly. On the basis of this framework, Yin et al. [67] and Zou et al. [74] added the optical flow map from the real scene to this frame. Yin et al. [67] explored an unsupervised GeoNet framework for simultaneously estimating depth map, optical flow and camera motion from sequential image frames using a deep CNN network. Similarly, Zou et al. [74] proposed a DF-Net framework which can simultaneously predict and estimate the single-view depth and optical flow. Feng et al. [16] proposed to apply a stacked of generative adversarial network layers to progressively extract lower layers to estimate ego-motion and higher layers to learn the spatial features. Although these networks displayed good depth estimation effects, all unsupervised approaches require extra sequential stereo images or videos and fail to recover correct scale information from the camera.

2.3 Visual-inerial measurements

There are applications targeting at the combination of the visual camera and Inertial Measurement Unit (IMU) at the same time, which can be classified into two categories. One category is filter-based and optimization-based fusion methods, which is similar to SLAM systems. Another category is loosely- and tightly-coupled fusion. However many applications incorporate the two categories to VIO systems. Classical filter-based and tightly-coupled approaches are proposed including MSCKF [34, 39] and ROVIO [6]. Although they incorporate image features into feature vectors, the major difference of both methods is using different filtering methods (e.g., MSCKF based on the multi-state-constraint Kalman filter and ROVIO based on iterated extended Kalman filter). Other

improved tightly-coupled approaches are also proposed [28, 52]. Using the same filter-based methods, SSF and MSF algorithms utilize loosely-coupled method to fuse features into IMU data after VO processing instead of directly adding image features to feature vectors [60]. For optimization-based methods, the tightly-coupled approaches have been more rapidly developed than loosely-coupled, such as OKVIS method [32]. This method jointly optimizes visual features and inertial error using a keyframe-based approach. Recently, VINS-Monocular camera is regarded as a fast, tightly-coupled, and nonlinear optimization-based method and applied to reach accurate VIO systems by fusing pre-integrated IMU [46].

3 Self-supervised 3D root system reconstruction learning network

Given unlabeled monocular color image sequences and the corresponding IMU information from the gyroscope and accelerometer sensor, our proposed framework is able to learn a representation to regress a 6-DoF camera poses and recover a 3D point cloud from multiple colors and depth images. An overview of the self-supervised learning framework is demonstrated in Fig. 1. During the training process, adjacent color images (i.e. I_{t-1}, I_t and I_{t+1}) from a monocular camera are fed into the pose estimation network to regress a 6-DoF ego-motion. Additional IMU measurement on the same time interval is fused together with the regressed poses directly from the network by a pose consistency constraint to leverage both benefits of these two sensors. Simultaneously, the depth estimator aims to estimate a depth map for every single input by continuously warping the source images to the reference images. The geometric constraints enable the network to self-supervise without additional priors or ground truth labels, and the fusion strategy for IMU and the regressed poses provide a theoretically support to recover an absolute scale and realize a scale-consistent estimation. The proposed network provides a novel learning-based SfM method to reconstruct the 3D models without external supervision. Details are further explained in the following sections.

3.1 Deep structure-from-motion

With any two consecutive frames as input, our deep SfM network is able to estimate their depth maps D, and reconstruct them to a complete point cloud. The relative 6 DoF poses P_{rel} between them are also able to be derived from the pose estimation network.

With the predicted depth maps and the relative camera motions, we can build a self-supervision constraint between the reference images I_{ref} and the warped images $I_{ref}{}'$ from the source images I_{src} using the predicted depth D and relative poses P_{rel}. Different from [67, 71], we extend the 3 local frames for optimization in the training process to a full sequence for each update, which realizes a full Bundle Adjustment (BA) optimization for joint camera poses and depth estimation problem. Then the proposed BA appearance loss is as:

$$L_{BA} = \sum_{i} \sum_{k} \sum_{j} \left\| I_i(\pi(\mathbf{P_{rel}},\ D_j \cdot \boldsymbol{p_j})) - I_{src-k}(\boldsymbol{p_j}) \right\|_1 \tag{1}$$

where $\mathbf{P_{rel}}$ is the relative camera motion from the reference image to the source frames. D_j corresponds to the depth value of a pixel p_j at the i_{th} reference image I_i. I_{src-k} is the k_{th} source image in each image group. The L1 loss guidance here is to measure the difference between each projected scene point of the reconstructed source image with the original

source image and minimize it. $[\mathbf{P_1}, \mathbf{P_1}, ..., \mathbf{P_N}]$ and $[D_1, D_2, ..., D_N]$ are the parameters that the network aims to optimize.

As L1 loss alone is not robust to the light illumination and contrast variation in real applications, we add the image structural similarity index to jointly evaluate two images in illuminance, contrast and structure. The modified bundle adjustment loss is a comprehensive expression of SSIM and L1 loss as:

$$L_{BA} = \sum_i \sum_k \sum_j (\lambda_1 \left\| I_i(\pi(\mathbf{P_{rel}},\ D_j \cdot \boldsymbol{p_j})) - I_{src-k}(\boldsymbol{p_j}) \right\|_1 + \lambda_2 \frac{1 - SSIM(I_i, I_{src-k})}{2}) \tag{2}$$

where $SSIM(I_i, I_{src-k})$ represents the element wise similarity between the warped image from each reference I_i and the source image I_{src}. We set λ_1=0.15 and λ_2 = 0.85 following [21, 22].

Apart from the BA appearance loss in pixel intensity, we further enforce a global geometrical consistency in 3D space for predicted depth maps. More specifically, the depth value at a specific pixel position p in frame I_1 should be always consistent with the depth at the same position in frame I_2, I_3, ... I_N, where N is the total number of frames. The 3D consistency loss is therefore defined as:

$$L_{depth} = \sum_k \sum_i \sum_j |D_i(p_j) - D_{src-k}(p_j)| \tag{3}$$

Here D_i is the warped depth map from the reference view to the source view, and D_{src-k} is the corresponding depth map of the input source image. By applying L1 loss to the difference of each source-reference image pair, the inconsistency of global depth value is able to be prevented.

To further prevent potential discontinuities of the texture-less root regions, an edge-aware smoothness is introduced as an additional penalty term to further prevent small divergent depth values and suppress the outliers. We propose to minimize depth-Laplacian of Gaussian (LoG) filter whose each element is weighted by the corresponding image gradient as follows:

$$L_{smooth} = w \left\| \frac{|\bigtriangledown^2 (D \otimes G)|}{\|\bigtriangledown I\|} \right\|_1 \times \frac{1}{\|D\|} \tag{4}$$

where $\bigtriangledown$ and $\bigtriangledown^2$ refer to the gradient and Laplacian operator respectively. G represents a 5×5 Gaussian kernel. D and I correspond to the predicted disparity map and input image respectively. w is the weight parameter and is set to 0.25 in our network. The motivation to scale the first term by dividing the mean disparity value is that the output can be normalized.

3.2 Fusion of structure-from-motion and IMU information

Monocular SfM algorithms exist an inherent limitation that the reconstruction can only be determined up to the unknown scale factor. The integration of the IMU sensor is able to recover a correct scale via the information of acceleration and rotation angles between two consecutive frames as shown in Fig. 2 and (5), where the V and T represent the motion speed in each axis and the corresponding motion distance, and a is the acceleration reading from the accelerator. The fusion of the SfM structure and the inertial measurement from IMU are utilized to estimate the real scale of the scene reconstruction model by constructing

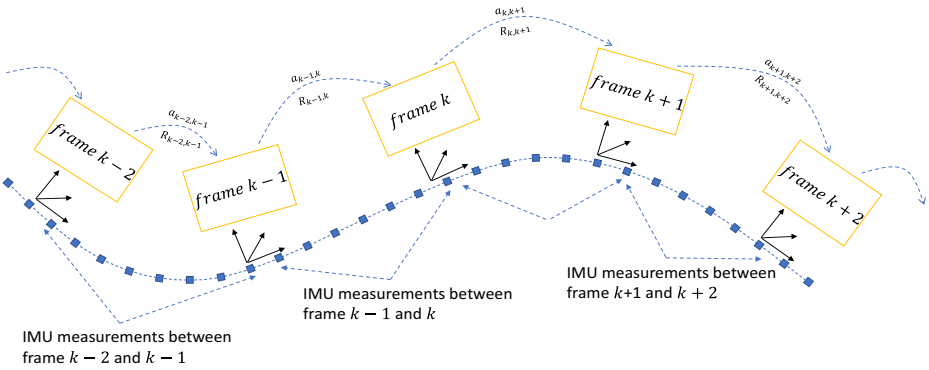

Fig. 2 Illustration of integration of IMU sensor and visual SfM reconstruction

a pose consistency loss to dynamically penalize the deviation of the pose from the pose estimation network and the IMU sensor to obtain a correct scale.

$$V = \int \vec{a}\, dt \quad \textbf{and} \quad T = \int (\int \vec{a}\, dt) dt \tag{5}$$

$$L_{scale} = \sum_{i=N} \sum_{t=x,y,z} |P^{imu}_{i,t} * P^{net}_{i,t} - P^{net}_{i,t} * P^{net}_{i,t}| \tag{6}$$

where N is the total number of frames. x, y, z are the three dimensions in translation vector. P^{imu} stands for the relative pose translation from IMU sensor and P^{net} stands for the estimated camera translation from the deep network.

We apply a depth fusion strategy to integrate multiple depths from our network at different viewpoints to a unified and complete point cloud. By continuously computing each depth map for one specific view and treating all other views as the reference, all depth maps are able to be fused into a complete point cloud. To prevent possible outliers and reduce the noise, our framework generates the best possible point cloud by a consistency examination via reprojecting each image to other N-1 views. If a depth value is consistent with the reference view within a threshold, the corresponding pixels will be preserved. Otherwise, those pixels with large reprojection errors will be removed. The illustration of this process is shown in the Fig. 3. Here, X is the common scene point corresponding to the X', X'', X''' and X'''' at different views in 2D image. The final point cloud is composed of all registered point clouds in 3D space.

Finally, our comprehensive loss constraints turn into a linear combination of: 1) BA appearance loss between each source image and warped reference images in a full sequence; 2) Depth consistency loss to enforce the global geometrical consistency in 3D space; 3) Edge-aware smoothness loss to prevent large discontinuities within each object; 4) Scale consistency loss to dynamically correct the predicted camera poses from the network via IMU measurement.

$$L = w_1 L_{BA} + w_2 L_{depth} + w_3 L_{smooth} + w_4 L_{scale} \tag{7}$$

where w_1, w_2, w_3 and w_4 are their associated weights for BA appearance loss, depth consistency loss, edge-aware smoothness loss, and scale consistency loss respectively.

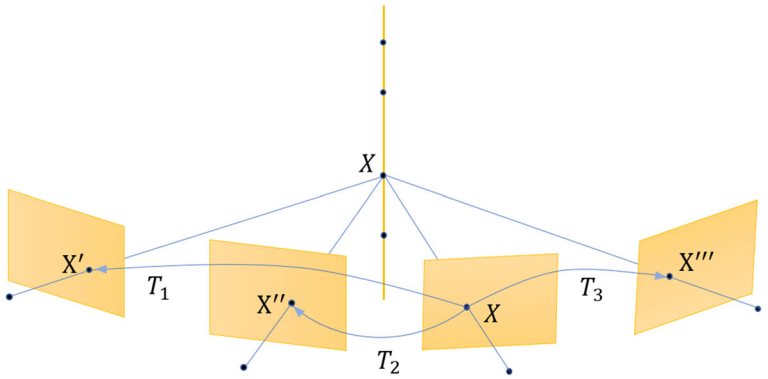

Fig. 3 Point cloud generation from multi-view setting

4 Experiments

In this section, we first describe the methods we used to collect real plant root dataset and creat a synthetic root dataset. We then introduce the configuration settings and the evaluation metrics to compare the proposed approach with other existing methods, and present the training and evaluation setup. Then, we compare both qualitative and quantitative results from the proposed network with other most recent methods. Without using ground truth 3D models for supervised training, we show that the proposed model achieves superior performance.

4.1 Dataset introduction

In this work, we verify the effectiveness of the proposed method on both real and synthetic datasets. First, we introduce the collected dataset on real plant roots. The dataset was captured by the visual camera of iPhone 7 with its embedded gyroscope and accelerometer. The plant root systems used to capture root images are those of one year old apple rootstocks generated from cuttings by a commercial nursery. The images of root system of apple rootstocks were taken/shown upside-down. The visual camera was used to record a series of image sequences, and the IMU sensor was used to export the corresponding rotation and translation information from consecutive frames. There are a total of 10000 images of 20 different plant root systems, each with around 500 images. We split 15 plant roots' images in the dataset for training and the rest five plant roots' images for testing. Each plant root is captured under a white background to prevent the possible reconstruction of background, with a rough coverage of 360 degrees' viewpoints and a resolution of 1920×1080. To synchronize the images from the camera and data from the IMU sensor, we first use mobile software to control them to start at almost the same time, and then utilize the returned timestamps to further make them consist in time. The frequencies for camera and IMU are set to be 30 Hz. The dataset is collected in a setting that roots are fixed to a stand without moving during data collection process. In applications involving moving objects, the target objects can be separated based on different motions of moving and stationary objects analyzed from visual tracking and IMU data analysis. The configuration setting of the equipment is shown in the Table 1.

Table 1 Plant root dataset was collected from one year old rootstocks of apples generated from cuttings using camera from iPhone 7, its embedded gyroscope and accelerometer sensor

Sensor name	iPhone 7 main camera and IMU
Pixel dimension	1920 × 1080
Length of Focus	28mm
Sensor Size	1/3" = 4.80 x 3.60 mm
Aperture	f/1.8

In addition to the real dataset, we also create a synthetic dataset based on the GameObject class in Unity [23]. By creating and controlling synthetic 3D models for each kind of plant root, we enjoy a similar setting with the real dataset. First, we are able to extract a sequence of root images by continuously rotating each 3D model and projecting them into the 2D image. Then by controlling the camera position and rotation, we are able to obtain the ego-motions which are similar to the egomotion data we retrieve from the IMU sensor. In addition to the two inputs, the 3D synthetic root model is further used as the ground truth to compare our 3D reconstruction model with other existing methods quantitatively. Sample images from our collected real and synthetic dataset are shown in Figs. 4 and 5.

4.2 Training configuration

During the training process, we first scale the image sequence from 1920 × 1080 to 512 × 256. For each iteration, each group of images together with the relative translation and rotation vectors from IMU sensor are input to the network. Each group of images are fed into the pose estimation network and depth prediction network simultaneously to learn representations of relative poses and depth maps. IMU information are used to dynamically correct the output from the pose estimation network to produce a complete 3D point cloud

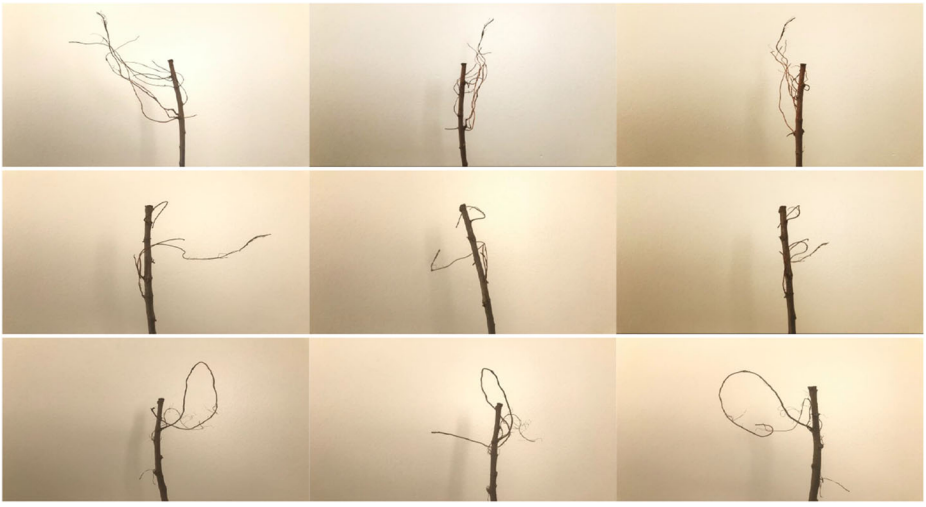

Fig. 4 Demo images from the real plant root systems of one year old apple rootstocks generated from cuttings by a commercial nursery. The images of root systems of apple rootstocks were taken/shown upside-down. Each row contains three images captured from different viewpoints

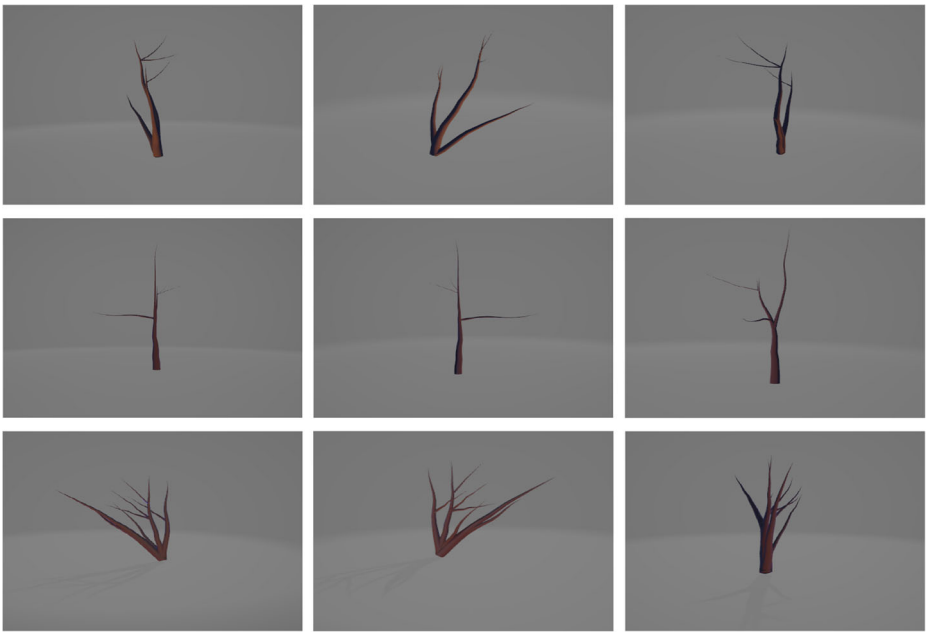

Fig. 5 The created synthetic root dataset. The dataset was rendered from 3D root models created by Unity 3D module. The first, second and third rows represent the projected images from different viewpoints for each 3D root model. With consecutive images from multiple views and accurate ego-motion relationships simulating IMU sensor as input, a completed 3D reconstructed model is able to be inferred

of the plant roots. Compared with the current setting applied in the experiment, IMU with lower drift and higher accuracy will help enhance root system reconstruction effect, but not significantly. IMU mainly accumulates motion estimation errors in long sequences. Our collected videos for each root system are short sequences (around 500 images under 30fps). The introduced error would substantially small in such short videos. Furthermore, instead of the absolute IMU poses calculated from the accumulated IMU motion, the relative motion between image frames is applied to the training and testing processes, especially in the application of the scale consistency loss during sensor fusion, which can reduce the errors in the training process and enhance the robustness in real-time applications.

The network is implemented with PyTorch and trained from scratch using Adam optimizer [30] with $\beta 1 = 0.9$ and $\beta 2 = 0.99$. The initial learning rate starts with 1e-4 and gradually decay to half of the original size for every 10 epochs in our framework. Rectified Linear Units (ReLU) [40] is utilized as the non-linear activation functions for convolutional layers. The weights of the depth estimation network and pose estimation network are initialized with Kaiming initialization [25] method with a batch size of 4 to easily model non-linearity of ReLUs. The whole framework is trained for 30 epochs to get good predictions.

4.3 Evaluation of the 3D root system reconstruction

A favorable reconstruction model usually maintains with reasonable and consistent shape as the input image, tight points, and fewer outliners or noise. To quantitatively examine the 3D

reconstruction performance of our method on synthetic datasets, we apply the following two prescribed metrics: mean distance and variance. Targeting at evaluating the performance of the generated 3D point cloud, the mean distance measures the mean value of the distance of each predicted point corresponding to a 3D point cloud of ground truth, which can truly reflect difference of our predicted reconstruction of root system compared with the ground truth. And the variance reflects the vibration of each point cloud compared with the ground truth point cloud. The smaller discrepancy in mean distance and variance indicates a better reconstruction performance compared with other methods.

4.4 Comparison with state-of-the-art methods

Qualitative results and analyses We first compared the results of the depth estimation network with one single image as input for testing in the real applications. From Fig. 6, it can be observed that our outputs present a reasonable estimation which show brighter for closer objects and darker for farther branches. Compared with our method, [22] suffers from the issue of miss-prediction in depth values of root and leaf, and discontinuity in the bottom part of each roots. Though [5] improves the miss-prediction regions in [22], it still contains the intense discontinuity at the root segment, which results in a large gap or stretch in the reconstructed 3D point cloud.

Then we compare with other recent 3D reconstruction approaches on the real collected dataset, as Fig. 7. The results demonstrate that our proposed model is able to generate a more accurate point cloud from images at different viewpoints. We observe many stretching and outlier pixels on the results of [22] because they are unable to further constrain the back regions of roots from just a single view. Though without too many stretching regions and outliers in [57] compared with [22], the method still suffer from severe deformation in shape and wrong prediction in multiple root branches. The result of the traditional offline SfM method [65] presents an overall correct shape, but at the same time, the reconstructed result demonstrates very sparse points in 3D space and lost extensive details such as thin and small branches because of its deficiency in extracting and matching sufficient features for the root structure with a rather similar appearance. Compared with the methods mentioned above,

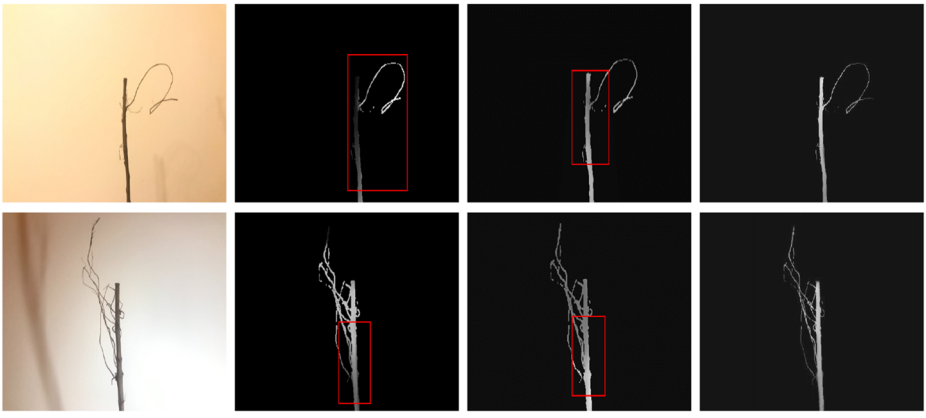

Fig. 6 Depth estimation results on single raw image as input. From left to right: raw input color image from the mobile device; depth estimation result from [22]; depth estimation result from [5]; our depth estimation results. Wrong predictions, faults and obvious discontinuities in the compared methods are marked with red rectangles

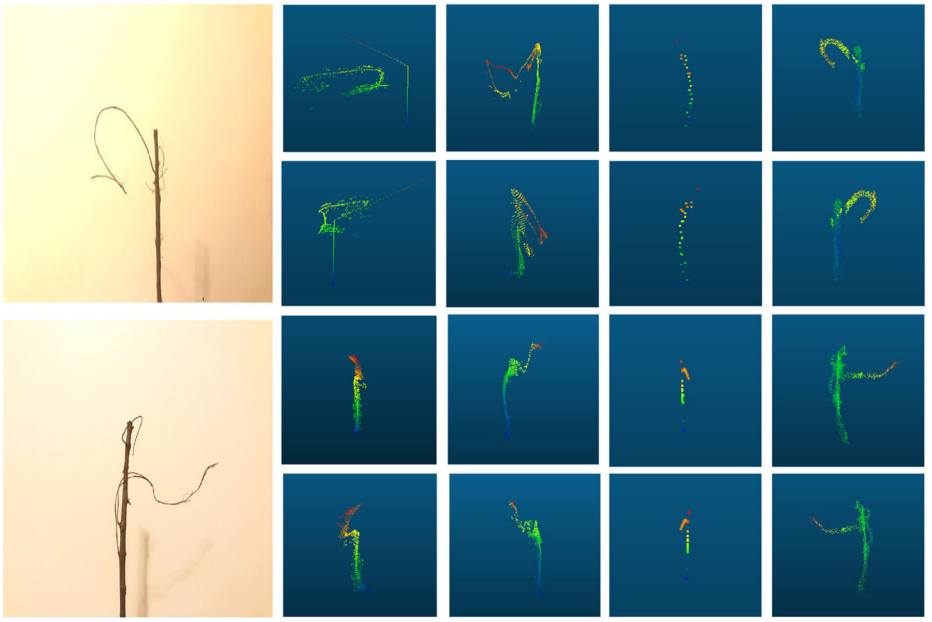

Fig. 7 Comparison of the reconstructed roots between our method and other state-of-the-art methods on real plant root dataset. The background is filtered out during the inference. From left to right: input image; results from Monodepth2 [22]; results from MVDepthNet [57]; results from classic Visual SfM [65]; results from our proposed method. Our reconstruction model can capture more accurate shape of plant root structure, and prevent many outliers and wrong-predictions

our proposed learning-based framework is able to eliminate incorrect depth mappings and generate an accurate 3D point cloud that enjoys the least wrong prediction and stretching effect because of our multi-view geometrical constraints and fusion strategy.

When comparing with the same methods in the synthetic dataset, almost every method shows a better reconstruction performance compared with that on the real dataset except for the classical SfM method [65], as shown in Fig. 8. The improvement for most methods could be explained from the high-quality rendered images of the created 3D model with a clear background. However, the clean and uniform images also result in a even more challenging situation for the classical SfM methods in feature extraction and matching, which can be observed from the bad performance in [65]. Compared with the results from [22] and [57], our method is able to produce the least distortion and stretching effect in shapes. Compared with [65], our framework is able to capture and recover more details in complex and thin branches, which are missing in the aforementioned methods.

Quantitative results and analysis Table 2 quantitatively compares the average distance and variance of the reconstructed point clouds generated from [5, 22, 57, 65] and the proposed method, compared with ground truth 3D point cloud from the synthetic 3D root structure models. To evaluate the effectiveness of different methods better, we take a combination of both the mean distance as well as variance to measure the quality of 3D reconstruction. We can see that the proposed method is capable to achieve smaller errors than other methods. We notice that the result from the proposed network with the robust loss constraints from multiple views and IMU sensors significantly outperforms the

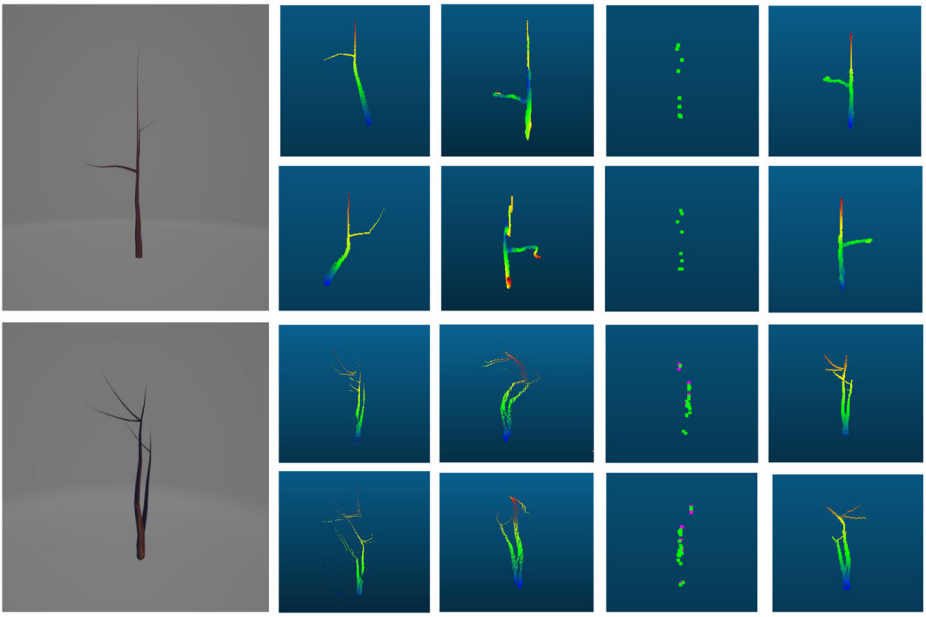

Fig. 8 Comparison of the reconstructed roots between our method and other state-of-the-art methods on the synthetic root dataset. The background is filtered out during the inference. From left to right: input image; results from Monodepth2 [22]; results from MVDepthNet [57]; results from classic Visual SfM [65]; results from our proposed method. Our reconstruction model is capable to recover the correct shape of root structure with the least distortion and wrong-predictions

variants from other recent classic and learning-based methods across all metrics. Benefiting from the correct scale recovered from IMU sensor, the result from the proposed full pipeline also witnesses an obvious improvement over the same network without it. Compared with [57], which also takes multiple view images as input, the proposed method realizes a 12.1 improvement in the mean error, and compared with [22], the presented result realizes a 77% and 81% lower mean distance and variance distribution respectively as shown in Table 2.

Table 2 Mean distance and variance to the ground truth of our method compared with other existing method taking single-view and multi-view images as input for scene reconstruction on the created synthetic root dataset. Results demonstrate that the mean distance and variance of the proposed method are smaller than other methods

Method	Mean distance	Variance
Monodepth2 [22]	32.63	13.58
MVDepthNet [57]	19.35	6.89
SC-SfMLearner [5]	24.41	7.91
SfM [65]	29.07	6.39
Ours w/o IMU	22.39	7.35
Ours	7.25	2.47

5 Conclusion

In this paper, we present a deep framework for Structure-from-Motion to generate a complete 3D root model from multiple images as input. In the unified framework, we enforce consistency of the scale of camera motion from IMU measurement, a geometric depth consistency and a bundle adjustment appearance consistency to jointly constrain each other across the whole framework. Our pipeline is able to be regarded as a learning-based realization of Bundle adjustment algorithm, which leverages both IMU information and multi-view geometric benefits. The experiments demonstrate a superior performance of our self-supervised method compared with other state-of-the-art deep learning based methods and offline classic SfM approach. The framework can be easily applied to mobile phone platforms that are accessible to general users.

References

1. Agarwal S, Furukawa Y, Snavely N, Simon I, Curless B, Seitz SM, Szeliski R (2011) Building rome in a day. Commun ACM 54(10):105–112
2. Agarwal S, Snavely N, Simon I, Seitz SM, Szeliski R (2009) Building rome in a day. In: IEEE international conference on computer vision, pp 72–79
3. Almalioglu Y, Saputra MRU, de Gusmao PP, Markham A, Trigoni N (2019) Ganvo: unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In: 2019 International conference on robotics and automation (ICRA), pp 5474–5480. IEEE
4. Beardsley P, Torr P, Zisserman A (1996) 3d model acquisition from extended image sequences. In: European conference on computer vision, pp 683–695. Springer
5. Bian JW, Li Z, Wang N, Zhan H, Shen C, Cheng MM, Reid I (2019) Unsupervised scale-consistent depth and ego-motion learning from monocular video. arXiv:1908.10553
6. Bloesch M, Burri M, Omari S, Hutter M, Siegwart R (2017) Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. The Int J of Rob Res 36(10):1053–1072
7. Bloesch M, Omari S, Hutter M, Siegwart R (2015) Robust visual inertial odometry using a direct ekf-based approach. In: 2015 IEEE/RSJ International conference on intelligent robots and systems (IROS), pp 298–304. IEEE
8. Chen J, Ngo CW (2016) Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM international conference on multimedia, pp 32–41
9. Chen JJ, Ngo CW, Chua TS (2017) Cross-modal recipe retrieval with rich food attributes. In: Proceedings of the 25th ACM international conference on multimedia, pp 1771–1779
10. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3d-r2n2: a unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision, pp 628–644. Springer
11. Cui H, Gao X, Shen S, Hu Z (2017) Hsfm: hybrid structure-from-motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1212–1221
12. Dellaert F, Seitz SM, Thorpe CE, Thrun S (2000) Structure from motion without correspondence. In: Proceedings IEEE conference on computer vision and pattern recognition. CVPR 2000 (Cat. No. PR00662), vol 2, pp 557–564. IEEE
13. Fan H, Su H, Guibas LJ (2017) A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 605–613
14. Farenzena M, Fusiello A, Gherardi R (2009) Structure-and-motion pipeline on a hierarchical cluster tree. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV workshops, pp 1489–1496. IEEE
15. Faugeras OD, Luong QT, Maybank SJ (1992) Camera self-calibration: theory and experiments. In: European conference on computer vision, pp 321–334. Springer
16. Feng T, Gu D (2019) Sganvo: unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. IEEE Rob Autom Lett 4(4):4431–4437
17. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395

18. Frahm JM, Fite-Georgel P, Gallup D, Johnson T, Raguram R, Wu C, Jen YH, Dunn E, Clipp B, Lazebnik S et al (2010) Building rome on a cloudless day. In: European conference on computer vision, pp 368–381. Springer
19. Garg R, BG VK, Carneiro G, Reid I (2016) Unsupervised cnn for single view depth estimation: geometry to the rescue. In: ECCV, pp 740–756
20. Gherardi R, Farenzena M, Fusiello A (2010) Improving the efficiency of hierarchical structure-and-motion. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 1594–1600. IEEE
21. Godard C, Mac Aodha O, Brostow GJ (2017) Unsupervised monocular depth estimation with left-right consistency. In: CVPR
22. Godard C, Mac Aodha O, Firman M, Brostow G (2019) Digging into self-supervised monocular depth estimation. ICCV
23. Haas JK (2014) A history of the unity game engine
24. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge University Press
25. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: ICCV, pp 1026–1034
26. Huang W, Liu H, Wan W (2020) Online initialization and extrinsic spatial-temporal calibration for monocular visual-inertial odometry. arXiv:2004.05534
27. Jiang N, Cui Z, Tan P (2013) A global linear method for camera pose registration. In: Proceedings of the IEEE international conference on computer vision, pp 481–488
28. Jones ES, Soatto S (2011) Visual-inertial navigation, mapping and localization: a scalable real-time causal approach. The Int J Rob Res 30(4):407–430
29. Khan M, Gemenet DC, Villordon A (2016) Root system architecture and abiotic stress tolerance: current knowledge in root and tuber crops. Front Plant Sci 7:1584
30. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
31. Leutenegger S, Furgale P, Rabaud V, Chli M, Konolige K, Siegwart R (2013) Keyframe-based visual-inertial slam using nonlinear optimization. Proceedings of robotis science and systems (RSS) 2013
32. Leutenegger S, Lynen S, Bosse M, Siegwart R, Furgale P (2015) Keyframe-based visual–inertial odometry using nonlinear optimization. The Int J Rob Res 34(3):314–334
33. Li K, Ma J, Li H, Han Y, Yue X, Chen Z, Yang J (2019) Discern depth under foul weather: estimate pm2.5 for depth inference. IEEE Trans Industr Inform
34. Li M, Mourikis AI (2013) High-precision, consistent ekf-based visual-inertial odometry. The Int J Rob Res 32(6):690–711
35. Li X, Hou Y, Wu Q, Wang P, Li W (2019) Dvonet: unsupervised monocular depth estimation and visual odometry. In: 2019 IEEE visual communications and image processing (VCIP), pp 1–4. IEEE
36. Ma J, Li K, Han Y, Du P, Yang J (2018) Image-based pm2. 5 estimation and its application on depth estimation. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1857–1861. IEEE
37. Moulon P, Monasse P, Marlet R (2012) Adaptive structure from motion with a contrario model estimation. In: Asian conference on computer vision, pp 257–270. Springer
38. Moulon P, Monasse P, Marlet R (2013) Global fusion of relative motions for robust, accurate and scalable structure from motion. In: Proceedings of the IEEE international conference on computer vision, pp 3248–3255
39. Mourikis AI, Roumeliotis SI (2007) A multi-state constraint kalman filter for vision-aided inertial navigation. In: Proceedings 2007 IEEE international conference on robotics and automation, pp 3565–3572. IEEE
40. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
41. Nath Kundu J, Krishna Uppala P, Pahuja A, Venkatesh Babu R (2018) Adadepth: unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2656–2665
42. Newcombe RA, Lovegrove SJ, Davison AJ (2011) Dtam: dense tracking and mapping in real-time. In: 2011 International conference on computer vision, pp 2320–2327. IEEE
43. Poggi M, Tosi F, Mattoccia S (2018) Learning monocular depth estimation with unsupervised trinocular assumptions. In: 2018 International conference on 3d vision (3DV), pp 324–333. IEEE
44. Pollefeys M, Koch R, Van Gool L (1999) Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. Int J Comput Vis 32(1):7–25
45. Pollefeys M, Nistér D, Frahm JM, Akbarzadeh A, Mordohai P, Clipp B, Engels C, Gallup D, Kim SJ, Merrell P et al (2008) Detailed real-time urban 3d reconstruction from video. Int J Comput Vis 78(2-3):143–167

46. Qin T, Li P, Shen S (2018) Vins-mono: a robust and versatile monocular visual-inertial state estimator. IEEE Trans Robot 34(4):1004–1020
47. Rogers ED, Benfey PN (2015) Regulation of plant root system architecture: implications for crop advancement. Curr Opin Biotechnol 32:93–98
48. Schonberger JL, Frahm JM (2016) Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4104–4113
49. Snavely N (2011) Scene reconstruction and visualization from internet photo collections: a survey. IPSJ Trans Comput Vis Appl 3:44–66
50. Snavely N, Seitz SM, Szeliski R (2006) Photo tourism: exploring photo collections in 3d. In: ACM Siggraph 2006 papers, pp 835–846
51. Sweeney C, Sattler T, Hollerer T, Turk M, Pollefeys M (2015) Optimizing the viewing graph for structure-from-motion. In: Proceedings of the IEEE international conference on computer vision, pp 801–809
52. Tanskanen P, Naegeli T, Pollefeys M, Hilliges O (2015) Semi-direct ekf-based monocular visual-inertial odometry. In: 2015 IEEE/RSJ International conference on intelligent robots and systems (IROS), pp 6073–6078. IEEE
53. Thrun S, Burgard W, Fox D (2005) Probabilistic robotics. 2005. Massachusetts Institute of Technology, USA
54. Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW (1999) Bundle adjustment—a modern synthesis. In: International workshop on vision algorithms, pp 298–372. Springer
55. Ummenhofer B, Zhou H, Uhrig J, Mayer N, Ilg E, Dosovitskiy A, Brox T (2017) Demon: depth and motion network for learning monocular stereo. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5038–5047
56. Wang D, Pan Q, Zhao C, Hu J, Liu L, Tian L (2016) Slam-based cooperative calibration for optical sensors array with gps/imu aided. In: 2016 International conference on unmanned aircraft systems (ICUAS), pp 615–623. IEEE
57. Wang K, Shen S (2018) Mvdepthnet: real-time multiview depth estimation neural network. In: 2018 International conference on 3d vision (3DV), pp 248–257. IEEE
58. Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG (2018) Pixel2mesh: generating 3d mesh models from single rgb images. In: Proceedings of the European conference on computer vision (ECCV), pp 52–67
59. Wang S, Clark R, Wen H, Trigoni N (2017) Deepvo: towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: 2017 IEEE International conference on robotics and automation (ICRA), pp 2043–2050. IEEE
60. Weiss SM (2012) Vision based navigation for micro helicopters. Ph.D. thesis, ETH Zurich
61. Wilson K, Snavely N (2014) Robust global translations with 1dsfm. In: European conference on computer vision, pp 61–75. Springer
62. Wu A, Han Y (2018) Multi-modal circulant fusion for video-to-language and backward. In: IJCAI, vol 3, p 8
63. Wu C (2013) Towards linear-time incremental structure from motion. In: 2013 International conference on 3d vision-3DV 2013, pp 127–134. IEEE
64. Wu C, Agarwal S, Curless B, Seitz SM (2011) Multicore bundle adjustment. In: CVPR, pp 3057–3064. IEEE
65. Wu C et al (2011) Visualsfm: a visual structure from motion system
66. Yang B, Wen H, Wang S, Clark R, Markham A, Trigoni N (2017) 3d object reconstruction from a single depth view with adversarial learning. In: Proceedings of the IEEE international conference on computer vision, pp 679–688
67. Yin Z, Shi J (2018) Geonet: unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1983–1992
68. You Y, Wang Y, Chao WL, Garg D, Pleiss G, Hariharan B, Campbell M, Weinberger KQ (2019) Pseudo-lidar++: accurate depth for 3d object detection in autonomous driving. arXiv:1906.06310
69. Zebedin L, Bauer J, Karner K, Bischof H (2008) Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. In: European conference on computer vision, pp 873–886. Springer
70. Zhan H, Garg R, Weerasekera CS, Li K, Agarwal H, Reid I (2018) Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: CVPR
71. Zhou T, Brown M, Snavely N, Lowe DG (2017) Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1851–1858
72. Zhu S, Shen T, Zhou L, Zhang R, Wang J, Fang T, Quan L (2017) Parallel structure from motion from local increment to global averaging. arXiv:1702.08601

73. Zhu S, Zhang R, Zhou L, Shen T, Fang T, Tan P, Quan L (2018) Very large-scale global sfm by distributed motion averaging. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4568–4577
74. Zou Y, Luo Z, Huang JB (2018) Df-net: unsupervised joint learning of depth and flow using cross-task consistency. In: Proceedings of the European conference on computer vision (ECCV), pp 36–53