Self-Supervised Convolutional Neural Networks for Plant Reconstruction Using Stereo Imagery

Yuanxin Xia[1], Pablo d'Angelo[1], Jiaojiao Tian[1], Friedrich Fraundorfer[1, 2], Peter Reinartz[1]

[1]Department of Photogrammetry and Image Analysis, Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Wessling, Germany

(Yuanxin.Xia, Pablo.Angelo, Jiaojiao.Tian, Peter.Reinartz)@dlr.de

[2]Institute of Computer Graphics and Vision, Graz University of

Technology (TU Graz), 8010 Graz, Austria

fraundorfer@icg.tugraz.at

Dense matching strategies combining convolutional neural networks and semi-global matching for plant reconstruction.

Abstract:

Stereo matching can provide complete and dense 3D reconstruction to study plant growth. Recently, high-quality stereo matching results were achieved combining semi-global matching with deep learning. However, due to a lack of suitable training data, this technique is not readily applicable for plant reconstruction. We propose a self-supervised MC-CNN scheme to calculate matching cost and test it for plant reconstruction. The MC-CNN network is re-trained using the initial matching results obtained from the standard MC-CNN weights. For the experiment, close-range photogrammetric imagery of an in-house plant is used. The results show that the performance of self-supervised MC-CNN is superior to the Census algorithm and comparable to MC-CNN trained by a LiDAR point cloud. Another experiment is performed using stereo imagery of a field beech tree. The proposed self-training strategy is tested and has proved capable of identifying the drought condition of trees from the reconstructed leaves.

1 Introduction

Forest management is an interdisciplinary topic involved in numerous fields such as environment, politics, economics, climate and ecology (Strigul, 2012). Remote sensing, as a technique to take measurements from a distance, is appropriate to assist forest management because it can observe the target with no need to approach it and provide time series data sets for constant monitoring. Spaceborne and airborne remote sensing instruments offer broad observation of trees to estimate the biomass, monitor the living condition, measure the forest canopy cover, etc. (Ahmed et al., 2014; Freeman et al., 2016; Wu et al., 2016). Some high-resolution stereo imaging sensors are capable of deriving detailed digital surface models to acquire geometric parameters of the forest, however, only some large scale properties such as forest canopy height can actually be estimated (Tian et al., 2017).

In order to obtain detailed information about the forest, single tree growth patterns should be observed. The size, shape, color and leaf distribution of individual trees are all important factors and worth measuring in detail so that the health situation of the tree and even the whole ecosystem can be better understood (Levin, 1999; Gatziolis et al., 2015). The terrestrial Light Detection and Ranging (LiDAR) technique can provide accurate and dense point clouds of trees to support the geometric survey for tree-level parameters estimation (Kankare et al., 2013; Tao et al., 2015). Nevertheless, the data acquisition can require considerable manpower and material resources and can even be dangerous in extreme terrain. In the past decade, dense matching using optical stereo images has been widely used for 3D reconstruction. Among the different techniques, Semi-Global Matching (SGM) has outperformed most existing approaches in accuracy and efficiency (especially in remote sensing), and is used in many applications, for example building reconstruction, digital surface model generation, robot navigation and driver assistance (Hirschmüller, 2011; Kuschk et al., 2014; Qin et al., 2015). However, the performance varies when different matching cost calculation approaches are adopted. Many local features (e.g. Census, Mutual Information) have been used for the matching cost calculation (Hirschmüller, 2008; Hirschmüller and Scharstein, 2009). But, tree leaf matching remains very difficult due to the lack of unique features, many occlusions and repetitive structure.

Convolutional Neural Networks (CNN) (LeCun et al., 1998) are a popular topic in computer vision and have been used to solve many vision problems. Recently, an algorithm computing Matching Cost based on CNN (MC-CNN) was proposed (Zbontar and LeCun, 2016) in which a net is trained with supervised learning based on pairs of small image patches with known true disparity. Combined with SGM, MC-CNN has proved to outperform most previous algorithms thanks to a good extraction of the local image features and a trained similarity measure to compare the extracted feature descriptors. However, the ground truth collection is always a bottleneck for deep neural network based algorithms, which require huge amount of labeled data to train the net (Krizhevsky et al., 2012; Knöbelreiter et al., 2018). Ground truth acquisition for tree reconstruction via LiDAR sensors is complicated by the long scanning time required for capturing a dense point cloud. Any tiny movement of the leaf or branch during the laser scanning will cause the scanned point cloud to be inconsistent with the images, which limits its use for further training and evaluation. Hence, in this paper we follow the work of (Knöbelreiter et al., 2018) and propose a dense matching strategy combining SGM and a self-trained MC-CNN for plant reconstruction.

This paper is organized as follows: The MC-CNN based dense matching and the proposed training schemes are described in Section 2. Section 3 describes an indoor and an outdoor experiment, which demonstrate the feasibility of the proposed self-training strategy. Conclusions are drawn and an outlook for future research is provided in Section 4.

2 Methodology

2.1 Dense Matching

Dense matching attempts at establishing correspondences between every pixel in the image pair (Scharstein and Szeliski, 2002). Together with the known camera orientations, a dense point cloud can be obtained. Most dense stereo matching algorithms consist of the following four steps: Firstly, a similarity measure between two potentially matching pixels is computed to evaluate the matching cost. Then as the matching cost can be ambiguous, costs are usually aggregated in a local neighborhood. Global stereo methods then apply regularization to the aggregated costs, while local methods simply select the correspondence with the lowest matching cost. SGM combines local and global methods by regularizing the aggregated costs before determining each correspondence. Afterwards for rectified stereo pairs, a disparity map containing the horizontal shifts between the images is obtained (Bolles et al., 1987; Okutomi and Kanade, 1993). Finally, subpixel interpolation, left-right consistency check and outlier filtering are applied by most stereo algorithms.

2.2 CNN

CNNs (LeCun et al., 1998) have been used to solve several vision problems such as classification (Krizhevsky et al., 2012), recognition (Lawrence et al., 1997), etc. It is basically a feed-forward artificial neural network constructed by a sequence of layers with learnable weights and biases. A volume of activations are transformed into another when going through the layers, and finally certain scores are obtained as output at the end of the network, e.g. class scores for classification. Four types of layers are frequently used: (a) convolutional layers, in which each neuron is related to a local region of the input; (b) pooling layers, used to downsample the previous volume; (c) rectified linear units applying an elementwise activation function; and (d) fully-connected layers, which calculate the output by connecting each neuron to all the neurons of the previous volume for high-level reasoning. The network can be trained to reach its best performance with a sufficient amount of training samples.

2.3 MC-CNN

CNNs provide a new possibility in dense matching (Luo et al., 2016; Zbontar and LeCun, 2016). Zbontar and LeCun (2016) proposed a dense stereo algorithm using a CNN based matching cost combined with SGM and additional post-processing steps, which outperformed most previous stereo matching algorithms. Therefore this algorithm is utilized as the main framework in this paper.

2.3.1 Data Term

A binary classification data set is constructed for training the net, based on either the KITTI (Geiger et al., 2013; Menze and Geiger, 2015) or the Middlebury (Scharstein and Szeliski, 2002, 2003; Scharstein and Pal, 2007; Hirschmüller and Scharstein, 2009; Scharstein et al., 2014) stereo data sets with available ground truth disparity maps. At each image location, a positive and a negative training example are extracted. The positive example is a pair of patches from the left and right image respectively with the central pixels projected from the same object point, while the negative example is from a pair of patches where this geometric condition is not satisfied.

Two network architectures are designed and trained on the extracted training examples. Both of them are siamese networks with two sub-networks sharing the same weights (Bromley et al., 1993). The first two sub-networks transform a pair of image patches into two feature vectors describing the structure of each patch. The siamese network consists of several convolutional layers, each of which is followed by a rectified linear unit. The second part of the network computes the similarity measure using the two feature vectors. The first architecture uses the dot product of the normalized feature vectors as similarity measure. Therefore, it has a lower runtime and is called fast architecture. The second architecture, shown in Figure 1 and named accurate architecture, learns the similarity measure during training. The outputs of the two subnets are concatenated and passed through a number of fully-connected layers with a rectified linear unit following each of them. At the end, there is one more fully-connected layer which uses the sigmoid nonlinearity to produce the similarity score. In this paper, the accurate architecture is adopted due to the high-quality demand of plant reconstruction.

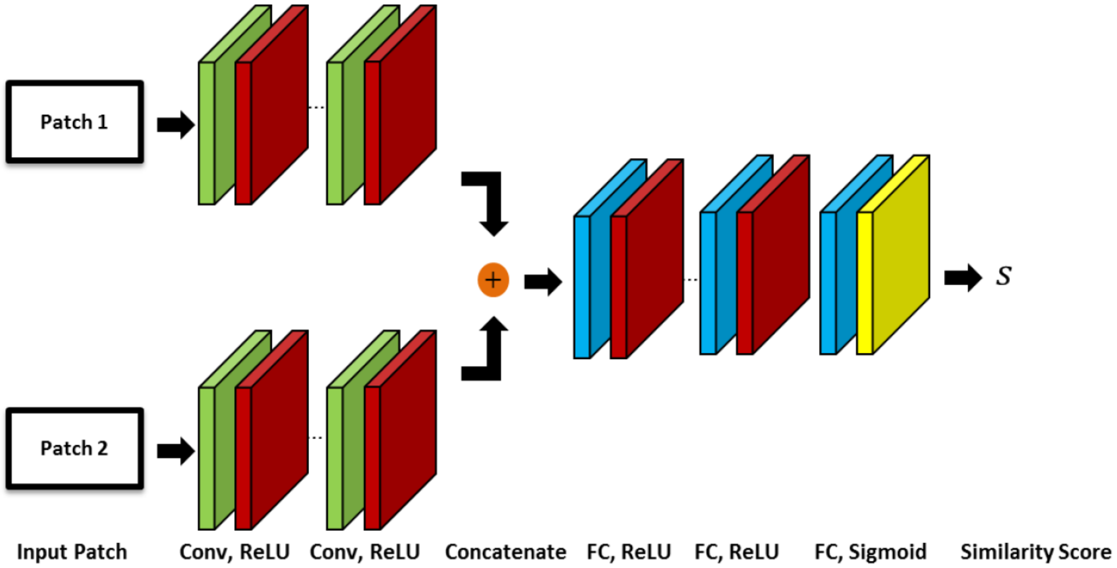


Figure 1. The accurate architecture computes the similarity score using fully connected network layers.

The binary cross-entropy loss used for training is defined as

$$l = t \cdot \log s + (1 - t) \cdot \log(1 - s), \tag{1}$$

in which l is the binary cross-entropy loss. s, the similarity score, represents the output of the net. The value of t depends on the category of the training example being used, which is equal to 1 for positive examples and 0 for negative examples. The hyperparameters include the number of convolutional layers in each subnet (5), the number of feature maps in each layer (112), the convolutional kernel size (3), the number of fully-connected layers (3), the corresponding number of units in each full-connected layer (384), and the input patch size (11×11). Zbontar and LeCun (2016) acquire the hyperparameters based on manual search and simple scripts to help automate the process, which are also applied in this paper.

2.3.2 Smoothness Term

SGM is used to regularize the disparity estimation using a piecewise constant smoothness term. SGM is a combination of local and global stereo matching methods (Hirschmüller, 2008), and approximates a global 2D smoothness term by summation of 1 dimensional smoothness constraints on 8 or 16 directions. For each direction, assuming the target pixel is at location p, the cost is computed as:

$$L_r(p,d) = C(p,d) + min(L_r(p-r,d), L_r(p-r,d-1) + P_1,$$

$$L_r(p-r,d+1) + P_1, min_i L_r(p-r,i) + P_2), \tag{2}$$

where $L_r(p,d)$ is the cost along the path traversed in direction r for the pixel p at disparity d and $C(p,d)$ is the matching cost. P_1 represents a penalty when the previous pixel has a disparity difference of 1. P_2 penalizes larger disparity differences. For each pixel p, $S(p,d) = \sum_r L_r(p,d)$ is computed and the disparity with the minimum S is selected.

SGM is selected as smoothness term due to its good performance and efficiency, its runtime is proportional to the reconstructed volume (d'Angelo and Reinartz, 2011; d'Angelo, 2016). $C(p,d)$ is calculated using MC-CNN and then aggregated based on Cross-Based Cost Aggregation (CBCA) (Mei et al., 2011; Zbontar and LeCun, 2016). It should be noticed that $S(p,d)$ undergoes CBCA once more before the final disparity determination.

2.3.3 Disparity Computation and Refinement

The disparity for each pixel is determined using the winner-takes-all strategy to generate a disparity map. Referring to Zbontar and LeCun (2016) and Mei et al. (2011), some post-processing steps are implemented to refine the quality of the disparity map, including interpolation, subpixel enhancement, a median filter, and a bilateral filter.

2.4 Training Details

As for the training, two schemes are designed, of which one utilizes the ground truth from a LiDAR scanner to construct training data, while the self-training scheme directly uses the dense matching results of MC-CNN, pre-trained on the Middlebury data sets, to re-train the network. The reason for the two schemes is to test how the performance of MC-CNN can be improved by self-training and training with ground truth, respectively.

2.4.1 LiDAR Training Scheme

Zbontar and LeCun (2016) provide several nets pre-trained on the KITTI 2012, KITTI 2015 and Middlebury data sets, respectively. The KITTI data sets focus on street views which do not fully match with our application. However, the Middlebury data focuses on static objects and the scenes exhibit a similar structure as our plant images, e.g. both concentrate on a certain target. Therefore, as one option we start from the pre-trained net on the Middlebury data sets and further train the net using the ground truth from LiDAR. In other words, we re-use the net pre-trained on the Middlebury data, and refine the network for plant reconstruction by further training. Thus the learning ability of the net for objects from a different category could also be tested.

As for the LiDAR scanning, a point cloud of the plant is generated to obtain the ground truth disparity map. As the image orientation and the LiDAR point cloud use different coordinate systems, a co-registration step is needed before the point cloud can be used. Besides, the main target is to test the performance of MC-CNN trained with different strategies for plant reconstruction and compare with a classic Census algorithm to demonstrate the effectiveness of MC-CNN. Hence as shown in Figure 2, we first generate two disparity maps based on SGM with Census and MC-CNN pre-trained on the Middlebury data sets. A pixel-wise average of both maps is computed and projected into 3D space to obtain a point cloud. Then, the point cloud from the laser scanner is registered to this newly generated point cloud. The ground truth disparity map is obtained by projecting the registered laser scanning point cloud onto the epipolar image planes. We use CloudCompare (Girardeau-Montaut et al., 2005) to roughly align the two point clouds first, by scale matching, rotation, translation and manual point pair picking alignment. After the rough alignment, some objects (in our case, leaves), which are reconstructed well by both dense matching and LiDAR, and aligned close to each other already, are selected for a further fine registration based on the Generalized Iterative Closest Point (GICP) method (Segal et al., 2009). GICP is more robust and performs better than the standard ICP without loss of efficiency. Afterwards, only well registered leaves are kept to generate the ground truth as described in detail by section 3.1.3.

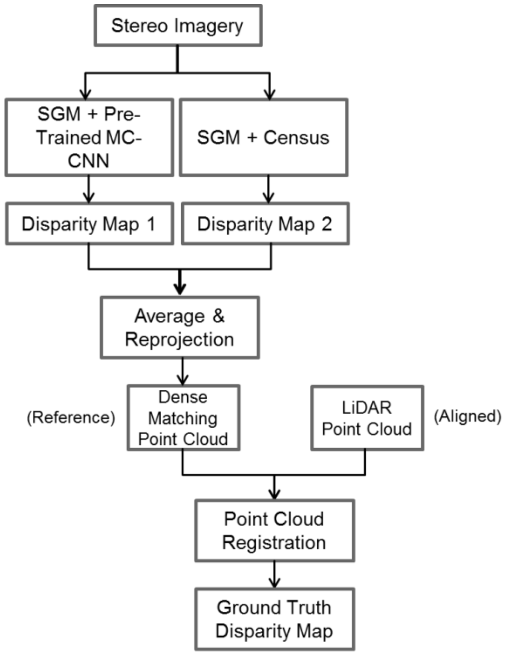


Figure 2. Flow chart for ground truth generation.

2.4.2 Self-Training Scheme

Huge amounts of data are available to meet the need of CNN for training. However in most cases, high performance is accomplished at the cost of substantial pre-processing workloads to label the training examples. Therefore, many self-supervised concepts have been proposed to avoid the time-consuming manual annotation (Joung et al., 2017; Zhou et al., 2017; Knöbelreiter et al., 2018). Joung et al. (2017) exploited the correspondence consistency between stereo images to pick samples during the training and guide the network to compute matching cost. Zhou et al. (2017) randomly initialized a network and adopted left-right consistency check to select suitable matching to train the net. Knöbelreiter et al. (2018) constructed the training data using a pre-trained version of their hybrid CNN-CRF model followed by a conservative consistency check to reject most outliers. Based on that, their self-supervised network is able to improve the completeness and accuracy of the stereo reconstruction results on aerial imagery.

Very high resolution LiDAR point clouds are very difficult and expensive to capture especially in an outdoor environment. In addition, it is almost impossible to obtain perfectly matching image and LiDAR data due to the long scanning time and changes in the plant shape due to wind and other effects. Therefore, instead of using LiDAR data, a self-training procedure is applicable even to scenarios where ground truth acquisition is difficult or impossible. We use the MC-CNN as described in section 2.3, pre-trained on Middlebury, to generate disparity maps used for self-training. A left-right consistency check with a threshold of 1 pixel is used to filter most outliers:

$$\left|d_p^L + d_q^R\right| \leq 1 \quad q = p - d_p^L, \tag{3}$$

where d_p^L is the disparity for pixel at location p in the disparity map regarding the left epipolar image as the master epipolar plane, while similarly d_q^R is calculated via dense matching regarding the right epipolar image as the master epipolar plane. Only pixels where left-right matching differs by less than 1 pixel are used as ground truth to further train MC-CNN.

3 Experiments

Two experiments demonstrate the feasibility of self-trained MC-CNN for plant reconstruction. The first experiment was carried out in an indoor laboratory environment. In this experiment, an 8-meter high tree standing in the atrium of a building was photographed from above. At the same time, a LiDAR point cloud was captured from a similar position. The second experiment investigated stereoscopic images from the crown of a beech tree growing in a typical European forest.

3.1 Experiment I

3.1.1 Data Set

The main objective of this work is the three-dimensional reconstruction of trees and their leaves in the forest. In order to minimize the influence of environmental conditions, the first experiment investigates an 8-meter high deciduous tree inside a building. A digital high-resolution handheld camera (NIKON D5500) equipped with an 18 mm lens is used to acquire images from a bridge over the crown of the tree. An exposure time of 1/20 seconds and an ISO speed rating of 400 was used. The acquired images are 4000 pixels in height and 6000 pixels in width. A stereo image pair with a baseline length of approximately 0.1 meters is taken from a distance of approximately 1 meter from the tree. Details about the image acquisition are available in Table 1. A Leica HDS7000 laser scanner is used to obtain a point cloud of the plant from a similar position. Capturing the point cloud with a point distance of 6.3 mm and a depth error of 0.4 mm RMS at a distance of 10 meters took about 10 minutes.

Table 1. The image acquisition parameters.

Camera model	NIKON D5500
Height	4000 pixels
Width	6000 pixels
Exposure time	1/20 sec
ISO speed rating	400

Focal length	18.0 mm
Object distance	≈ 1 m
GSD	0.02 cm/pixel
Baseline length	≈ 0.1 m

3.1.2 3D Reconstruction

The proposed dense matching approach requires epipolar images, where corresponding pixels are located on the same image row. MicMac (Rosu et al., 2015) was utilized for camera calibration, relative orientation and epipolar image rectification. The epipolar images generated based on the stereo pair mentioned above are shown in Figure 3.

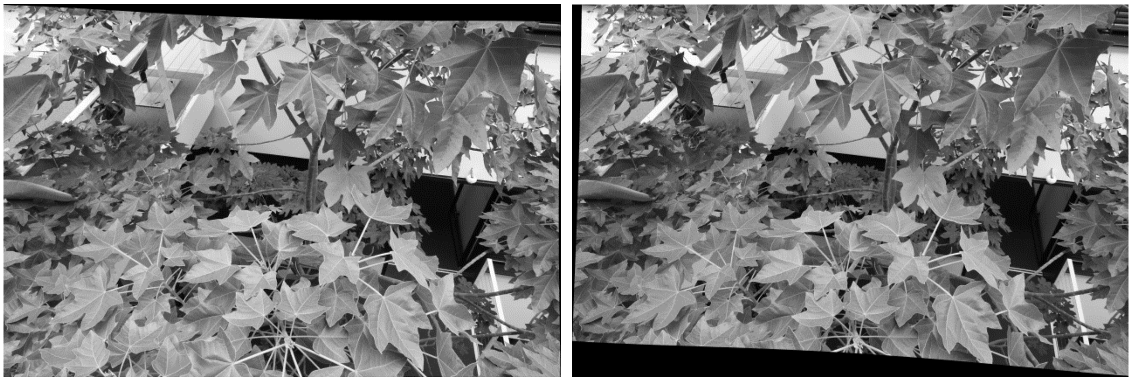


Figure 3. The epipolar image pair for dense matching.

Disparity maps have been calculated using the method described in sections 2.2 and 2.3 using 4 different matching costs:

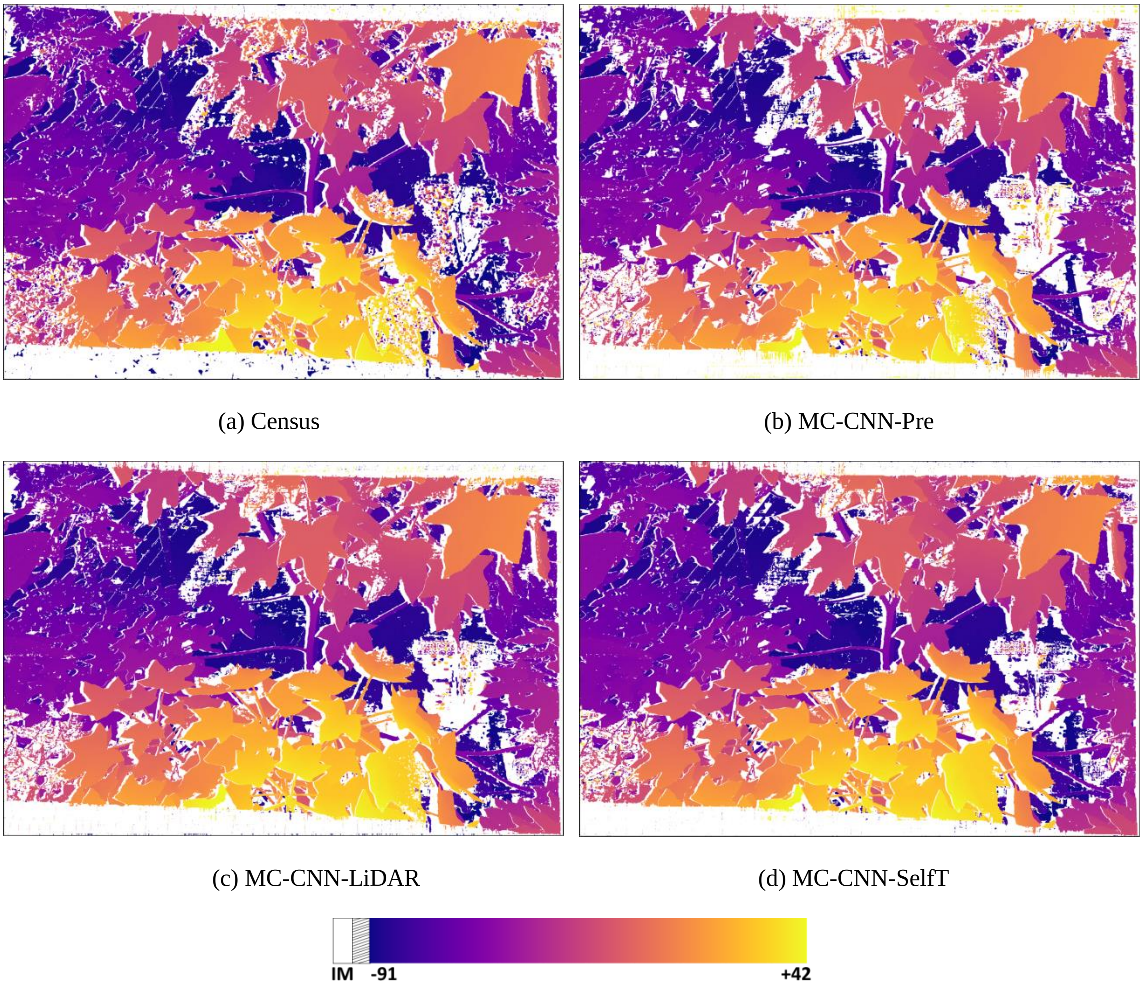
Census: Using only Census as matching cost;

MC-CNN-Pre: Using MC-CNN matching cost pre-trained on the Middlebury data sets;

MC-CNN-LiDAR: Using MC-CNN further trained on the LiDAR ground truth for matching cost, as described in section 2.4.1;

MC-CNN-SelfT: Using MC-CNN further trained using the disparity maps of MC-CNN-Pre, as described in section 2.4.2.

After the processing as described in section 2.3 and applying the left-right consistency check as described in section 2.4.2, the generated disparity maps for the epipolar image pair in Figure 3 are shown in Figure 4. For pixels with valid matching, the calculated disparity values from -91 to +42 are represented by the color from blue to yellow accordingly.



(a) Census

(b) MC-CNN-Pre

(c) MC-CNN-LiDAR

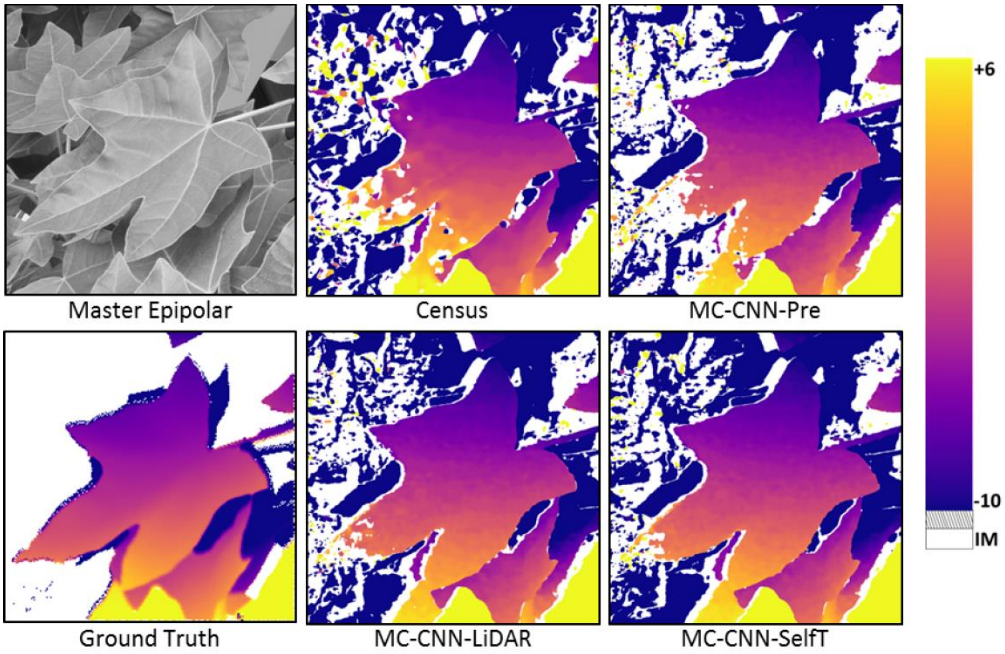
(d) MC-CNN-SelfT

Figure 4. The disparity maps generated based on SGM with different strategies for matching cost. Inconsistent matching (IM) is represented by the color white.

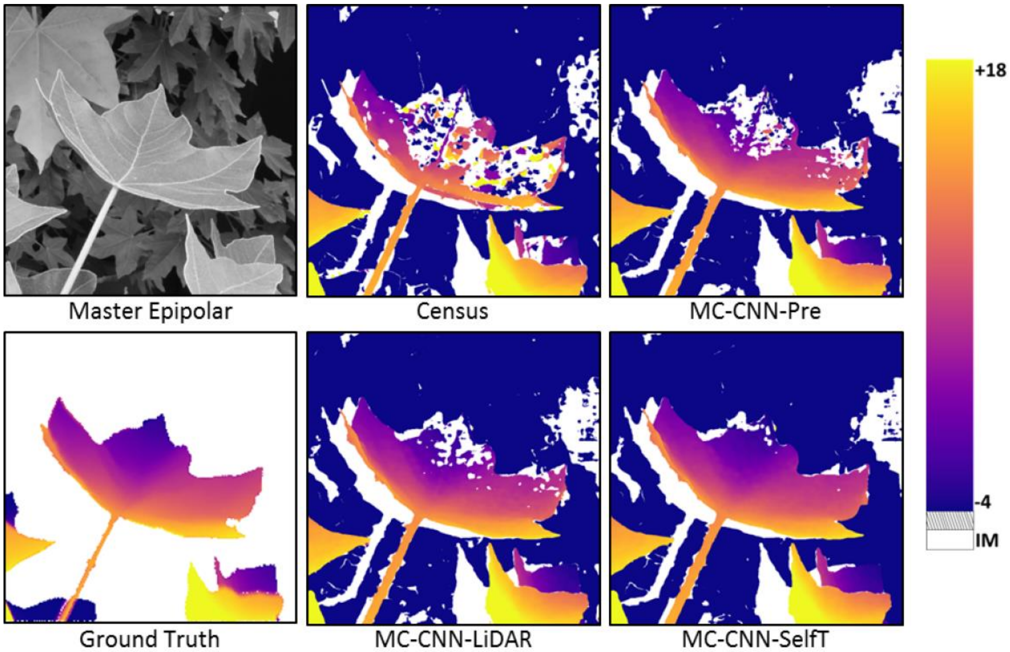
3.1.3 Evaluation and Discussion

Training and evaluation of the different methods is hampered by systematic differences between LiDAR and stereo pairs. Due to the automatic air conditioning of the building there were small movements of the branches and leaves during LiDAR recording which took around 10 minutes. These led to slightly different leaf positions between LiDAR and stereo images. During the generation of the ground truth disparity map, some errors are included unavoidably when picking up point pairs to align the point clouds initially. The fine registration with GICP can improve the co-registration but errors still exist. Due to these problems, the point cloud registration is not perfect which influences the use of the ground truth disparity map generated from the LiDAR data. This is also the reason that we determine to only focus on some selected leaves after rough alignment to do GICP, as mentioned in section 2.4.1. Afterwards the relatively well registered leaves by GICP, that visually show merely small shift between the point clouds, are utilized for training and evaluation of the methods, which alleviates the problem mentioned above. This is in accordance with our application, as the shape of the leaves is the major indicator of plant health. Compared with images from the Middlebury data sets with sizes of around 300×200 to 3000×2000 pixels, our images are larger (6000×4000 pixels), and the masked leaves can still provide a good amount of application specific training data. Thus, we use 13 well registered leaves together with Jadeplant and Sword1 data (containing a plant, belonging to the Middlebury data sets 2014) as training data. The reason for adding the Middlebury data into the newly generated data sets is to increase the amount of training data from limited selected leaves.

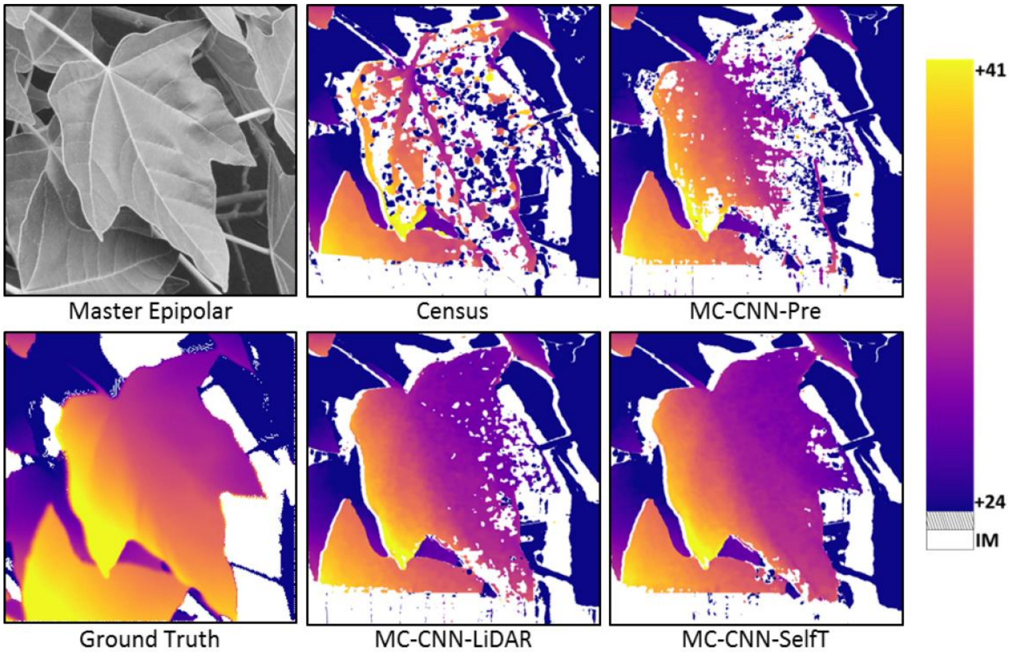
A visual comparison of the results in Figure 4 shows that the tree was well reconstructed by all matching schemes. The results of five independent leaves not used during training on the LiDAR ground truth are shown in Figure 5. While most parts of the leaves are well reconstructed, some differences in completeness and amount of outliers are visible.



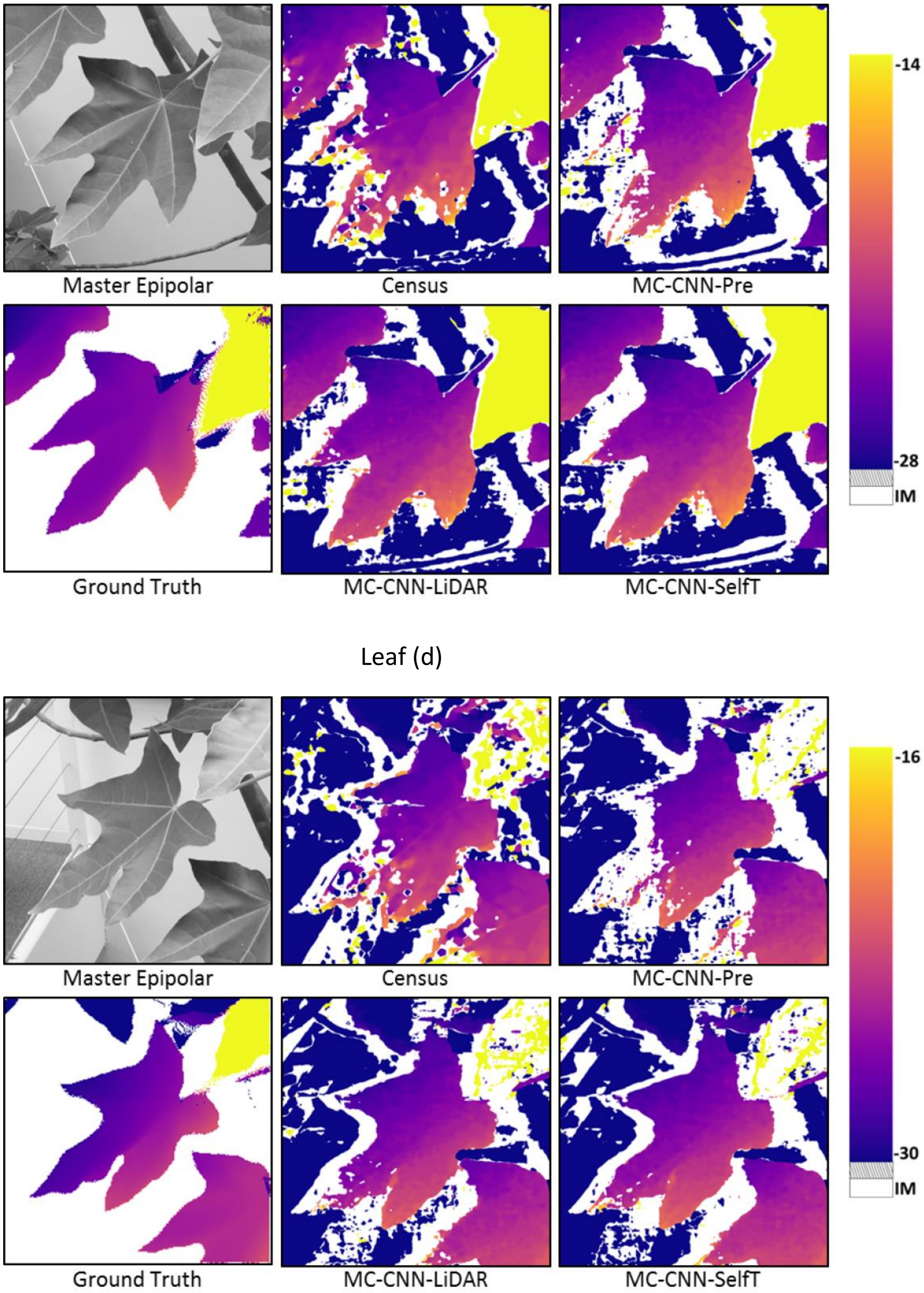
Leaf (a)


Master Epipolar
Census
MC-CNN-Pre
Ground Truth
MC-CNN-LiDAR
MC-CNN-SelfT
+18
-4
IM

Leaf (b)


Master Epipolar
Census
MC-CNN-Pre
Ground Truth
MC-CNN-LiDAR
MC-CNN-SelfT
+41
+24
IM

Leaf (c)



Leaf (d)

Leaf (e)

Figure 5. The reconstruction details of several selected leaves. From left to right in each subset: the first row includes the master epipolar image and dense matching results for

Census and MC-CNN-Pre. The second row includes the ground truth and dense matching results for MC-CNN-LiDAR and MC-CNN-SelfT. In order to enhance the contrast of the disparity within each single leaf, we have used a different colorbar for each leaf. Pixels invalidated by the left-right check are shown in white.

From a visual inspection, it is found that the disparity values obtained by all four strategies match with the ground truth. With Census as matching cost, the main shape of the leaf is reconstructed but with considerable noise and low completeness. MC-CNN-Pre results in low completeness, cf. leaf (e), but shows less noise. However when fed with specific data for further training, MC-CNN-LiDAR and MC-CNN-SelfT achieve higher reconstruction completeness. MC-CNN-SelfT results in a slightly better leaf reconstruction than MC-CNN-LiDAR and fewer gaps. We would like to point out two reasons for this behavior: Firstly, in self-training more training samples are available for the net to develop the ability to learn new feature and calculate the similarity score. In Figure 4, it can be seen that all leaves are reconstructed or partially reconstructed in MC-CNN-Pre. Hence, the further trained MC-CNN can learn from each single leaf during the training and recover more area. Besides the rigid left-right consistency check, applied to the dense matching results of MC-CNN-Pre to construct training samples, guarantees a reasonable training procedure for MC-CNN-SelfT.

A quantitative evaluation is performed by comparing the generated disparity maps with the disparity maps obtained from LiDAR. The leaves (a) – (e) shown above are used for comparison. Firstly, the disparity difference D_p is calculated as below in units of pixels:

$$D_p = d_p - d_p^G \quad p \in N_p, \tag{4}$$

where d_p denotes the disparity value of a pixel at location p calculated using one of the four dense matching schemes. d_p^G is the corresponding ground truth disparity value. N_p is the set of pixels where both dense matching and ground truth provide disparity values. The mean (D_{mean}), median (D_{median}), standard deviation (D_{STD}) and median absolute deviation (D_{MAD}) of the disparity differences are computed for comparison.

$$D_{mean} = mean(D_p) \tag{5}$$

$$D_{median} = median(D_p) \tag{6}$$

$$D_{STD} = \sqrt{mean(\,(D_p - D_{mean})^2\,)} \tag{7}$$

$$D_{MAD} = median\left(\left|D_p - D_{median}\right|\right). \tag{8}$$

The results are reported in Tables 2 to 5.

Table 2. Mean of the disparity difference between dense matching and ground truth.

	D_{mean} (pixels)			
leaf	Census	MC-CNN-Pre	MC-CNN-LiDAR	MC-CNN-SelfT
(a)	0.28	-0.23	**0.05**	0.17
(b)	-6.78	-4.96	-2.32	**-1.88**
(c)	-13.88	-14.32	-3.73	**-3.13**
(d)	**0.35**	0.72	0.50	0.64
(e)	-0.15	**0.14**	0.30	0.46

Table 3. Median of the disparity difference between dense matching and ground truth.

	D_{median} (pixels)			
leaf	Census	MC-CNN-Pre	MC-CNN-LiDAR	MC-CNN-SelfT
(a)	0.11	-0.11	-0.10	**-0.00**
(b)	-1.78	-1.72	-2.02	**-1.57**
(c)	-3.91	-3.30	-3.54	**-3.12**
(d)	**0.32**	0.48	0.40	0.57
(e)	**0.06**	0.29	0.28	0.40

Table 4. STD of the disparity difference between dense matching and ground truth.

	D_{STD} (pixels)			
leaf	Census	MC-CNN-Pre	MC-CNN-LiDAR	MC-CNN-SelfT
(a)	4.49	4.48	**2.37**	2.76
(b)	19.61	15.02	1.29	**1.28**
(c)	25.53	30.65	7.86	**6.38**
(d)	2.73	3.16	**1.06**	1.13
(e)	5.35	2.84	**0.70**	0.86

Table 5. MAD of the disparity difference between dense matching and ground truth.

	D_{MAD} (pixels)			
leaf	Census	MC-CNN-Pre	MC-CNN-LiDAR	MC-CNN-SelfT
(a)	0.76	**0.57**	**0.57**	0.63
(b)	3.03	0.51	0.42	**0.40**
(c)	3.49	0.64	**0.63**	**0.63**
(d)	0.73	0.67	**0.60**	0.65
(e)	0.50	0.46	**0.43**	0.51

By comparing the results in Table 2 and Table 3, it can be observed that the median is as expected more robust to outliers than the mean (e.g. for leaf (c), all the D_{median} are around 3 pixels). Leaf (b) and (c) show a relatively large systematic disparity difference. This can be attributed to the systematic error caused by the shape change and imperfect point cloud registration of the ground truth disparity map.

The D_{STD} values in Table 4 show the robustness of MC-CNN-LiDAR and MC-CNN-SelfT, as they exhibit much lower D_{STD} than Census and MC-CNN-Pre.

D_{MAD} has been widely used for depth map evaluation, as it is more robust to outliers than D_{STD}. The disparity map generated from Census has a relatively high D_{MAD} for the leaves (b) and (c). This is due to the large amount of noise in the Census results, as visible in Figure 5.

In addition to the pixel-based direct comparison, the reconstruction completeness and the percentage of the accurately measured pixels are calculated. The reconstruction completeness is calculated using the formula (9).

$$Cpl = \frac{n_{DM/G}}{n_G} \times 100\%, \tag{9}$$

where n_G denotes the number of pixels with a valid disparity value provided by the ground truth in each leaf. $n_{DM/G}$ denotes the number of pixels where both dense matching and ground truth provide disparity values. Thus the completeness Cpl will be the percentage of pixels in ground truth which are reconstructed by the dense matching as well.

However due to the systematic error, the disparity difference D_p between dense matching and ground truth cannot be directly utilized for evaluation. Therefore, we remove the systematic disparity shift for each leaf before computing the percentage of accurate pixels.

$$Acc = \frac{n_{pass}}{n_G} \times 100\% \tag{10}$$

$$n_{pass} = the\ \#\ of\ pixels \quad if: |D_p - D_{median_{mean}}| \leq \varepsilon \tag{11}$$

$$D_{median_{mean}} = mean(D_{median_{scheme\ i}}) \quad i \in \{1, 2, 3, 4\}, \tag{12}$$

where $D_{median_{mean}}$ is the mean of D_{median} calculated using each of the four matching schemes for each leaf. n_{pass} counts the number of pixels with the deviation below ε, a pre-defined threshold to evaluate the corresponding accuracy. In this paper, ε is set as 0.5 and 1 pixel respectively for the test. The results are shown in Table 6.

Table 6. Evaluation of reconstruction completeness and accuracy for each dense matching scheme.

Algorithm	(a)			(b)			(c)			(d)			(e)		
	Cpl	Acc		Cpl	Acc		Cpl	Acc		Cpl	Acc		Cpl	Acc	
		0.5 p	1 p		0.5 p	1 p		0.5 p	1 p		0.5 p	1 p		0.5 p	1 p
Census	92.0	31.8	57.0	63.0	14.8	23.9	49.7	7.6	14.0	92.0	36.4	56.9	89.7	43.3	71.0
MC-CNN-Pre	91.1	42.1	67.3	82.0	39.0	62.5	59.8	23.6	37.0	91.5	37.6	63.3	85.0	45.6	72.9
MC-CNN-LiDAR	96.9	**43.8**	**72.1**	89.2	**51.9**	70.7	86.4	34.5	60.5	**99.4**	**44.3**	**69.4**	97.1	**55.6**	**82.5**
MC-CNN-SelfT	**97.9**	41.0	67.0	**98.6**	51.0	**81.4**	**95.7**	**39.7**	**62.2**	**99.4**	41.9	67.8	**99.5**	47.9	77.4

MC-CNN-SelfT consistently obtains a slightly higher completeness than MC-CNN-LiDAR, while MC-CNN-LiDAR obtains slightly higher accuracy values for most leaves, except for leaves (b) and (c), where MC-CNN-SelfT shows significantly better completeness and 1 pixel accuracy values. Both re-trained methods consistently outperform Census and MC-CNN-Pre. This shows that especially MC-CNN-SelfT, which does not require additional LiDAR ground truth data, is a good approach for significantly improving the leaf reconstruction.

In this experiment, MC-CNN-LiDAR is handicapped due to imperfect ground truth, leading to disadvantages compared to the MC-CNN-SelfT method. We therefore assume that the scores for MC-CNN-LiDAR could be improved slightly by using a perfectly registered ground truth. However due to different registration errors for each leaf (cf. Table 3), the LiDAR trained network is not able to learn and correct for a systematic error between the LiDAR point cloud and the image data. We thus believe that the evaluation does not favor a specific method.

3.2 Experiment II

This work was performed as part of a project aiming at detecting the physiological and morphological status of trees under drought stress and studying the adaptation of forest areas to climate change. A major part of the project focuses on constructing a detailed and accurate 3D model of tree leaves in order to monitor the shape change when facing drought.

For this purpose, two nadir-viewing cameras are mounted on a crane system for stereo measurement. When the system is lifted above the trees, a stereo image pair of the tree crowns can be obtained. In order to test the feasibility of the stereo method described in this paper, a stereo image pair above a beech tree subject to slightly artificial drought stress is collected. Some information about the images and the camera setting is shown in Table 7.

Table 7. Details about the image acquisition.

Camera model	SONY ILCE-5100
Height	4000 pixels
Width	6000 pixels
Exposure time	1/60 sec
ISO speed rating	125
Focal length	19.0 mm
Object distance	≈ 3 m
GSD	0.06 cm/pixel
Baseline length	≈ 0.25 m
Acquisition date	June 19th, 2018

The corresponding epipolar image pair is shown in Figure 6. In this experiment, no LiDAR data is available, thus only Census, MC-CNN-Pre and MC-CNN-SelfT can be applied. The disparity map computed using MC-CNN-SelfT is shown in Figure 7.

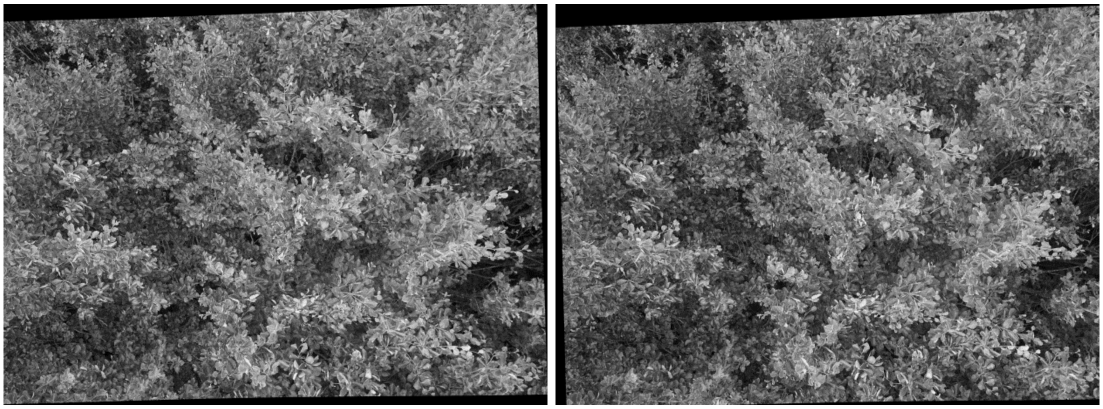


Figure 6. An epipolar image pair from the test region of our project.

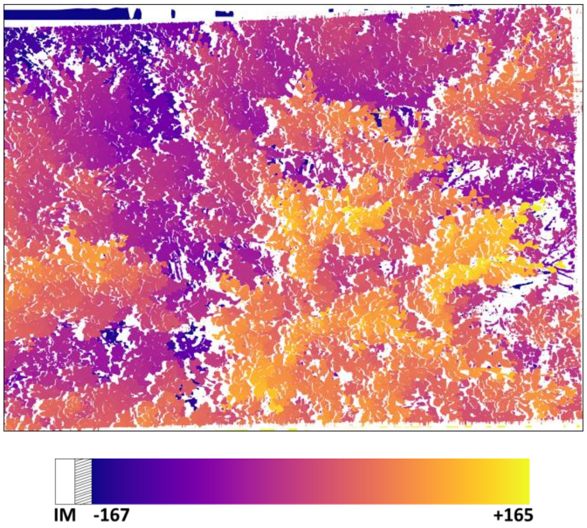
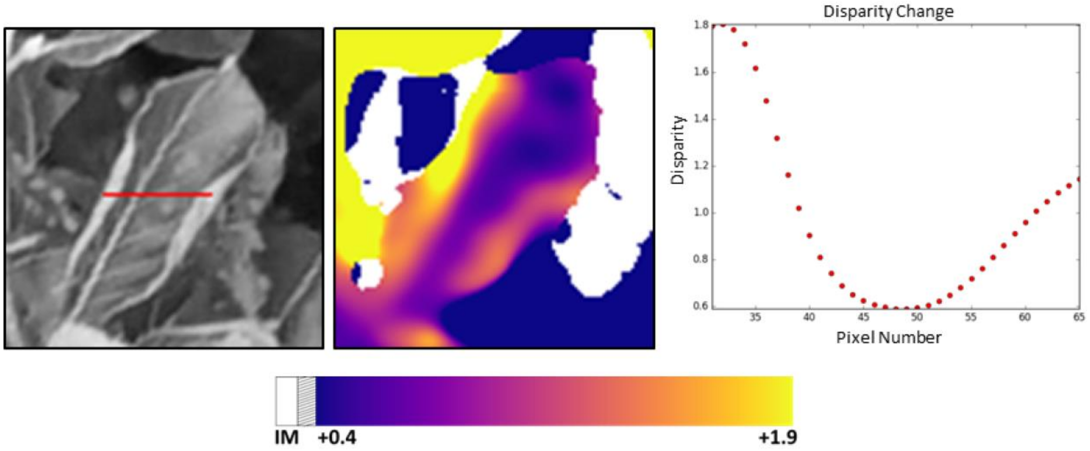
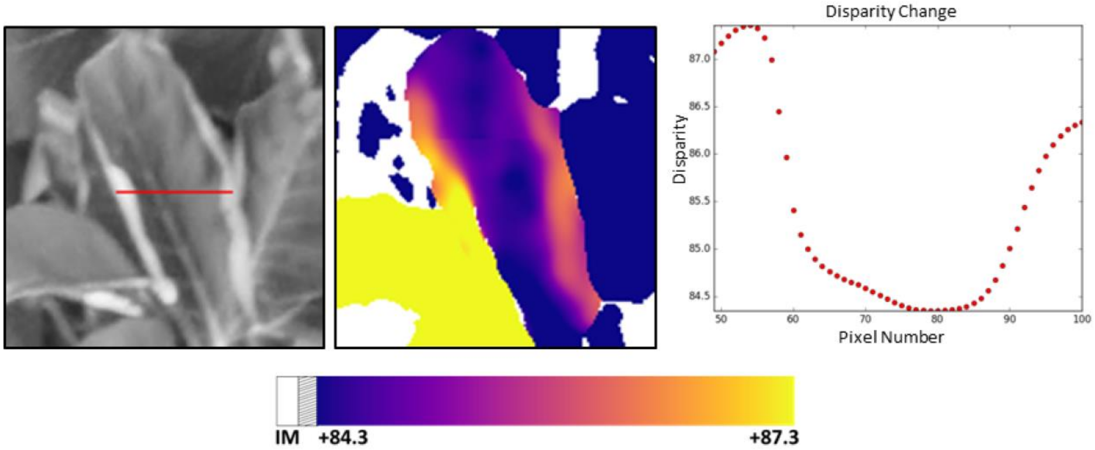


Figure 7. The disparity map generated using self-trained MC-CNN. Inconsistent matching is represented by the color white.

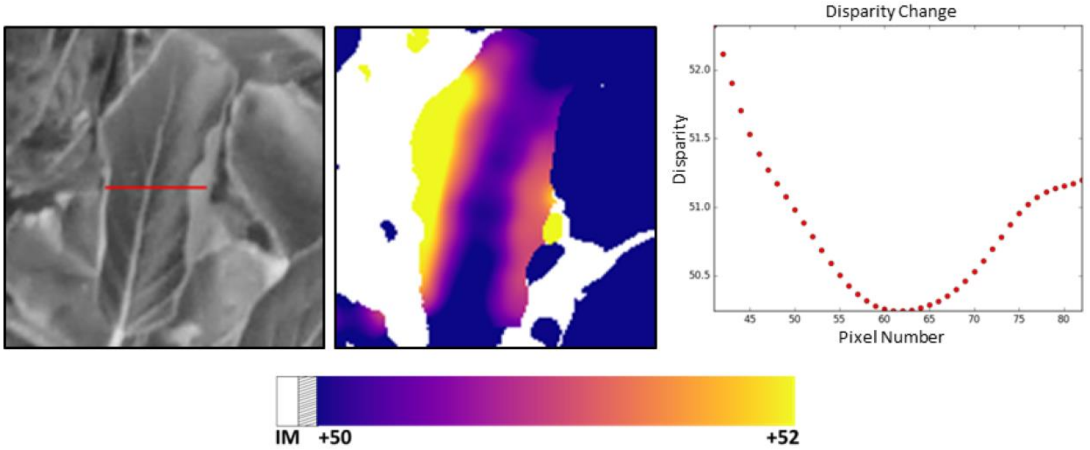
Figure 6 shows that the large beech tree crown is much more complex, and has much smaller leaves than the indoor tree used in the first experiment. The slight drought stress leads to multiple different leaf shapes. Under the hypothesis that curved leaves are an indicator for drought stress, the stereo method should enable a clear separation of planar and curved leaves. The generated disparity map provides a dense reconstruction of the tree crown, and individual leaves are separable. The reconstruction completeness for MC-CNN-Pre and MC-CNN-SelfT, are 76.0% and 78.7%, respectively. Due to the lack of ground truth, the value is computed as the ratio of pixel passing the left-right check to the number of valid pixels in the rectified image. Some leaves under drought stress are selected for visual comparison. As shown in Figure 8, the curled shape of the leaves is clearly visible in the disparity image and the profile plot.


Disparity Change
Disparity
Pixel Number
1.8
1.6
1.4
1.2
1.0
0.8
0.6
35
40
45
50
55
60
65
IM
+0.4
+1.9

(1)


Disparity Change
Disparity
Pixel Number
87.0
86.5
86.0
85.5
85.0
84.5
50
60
70
80
90
100
IM
+84.3
+87.3

(2)


Disparity Change
Disparity
Pixel Number
52.0
51.5
51.0
50.5
45
50
55
60
65
70
75
80
IM
+50
+52

(3)

Figure 8. Leaves under drought stress. From left to right in each subset: the master epipolar image, the disparity map of the self-trained MC-CNN matching scheme, and the disparity profile along the red line. The color represents the disparity. From blue to yellow, the targets get closer to the camera. Pixels with inconsistent matching are shown in white color.

It can be found that all the profiles are roughly U shaped, similar to the true shape of the leaves.

4 Conclusion

Plant reconstruction from stereo imagery is difficult due to the complexity of leaves which exhibit similar shape and intensity information. Hence the matching cost computation should be accurate to adequately represent the similarity between patches as the basis for the final disparity computation. SGM combined with MC-CNN has proved to outperform most previous algorithms; however, in practice it is extremely difficult to capture a large amount of high-quality training data. In this paper, a self-trained MC-CNN without the use of ground truth is tested to reconstruct the plant. Based on the dense matching results of MC-CNN pre-trained on the Middlebury data sets, a rigid left-right consistency check is applied to limit the outliers and the filtered results are utilized to further train the net. The reconstructed plant shows superior performance for the self-trained version than for the pre-trained one and the classic Census algorithm. Compared with MC-CNN further trained using the ground truth from LiDAR, the self-trained net behaves slightly worse in accuracy but better in reconstruction completeness. The self-training strategy of MC-CNN is also applied to the stereo imagery of a natural forest tree under drought condition. The resultant disparity map is capable of showing the deformation of leaves, which highlights the possibility of the self-trained MC-CNN to monitor the tree health situation.

In future research, more approaches will be tested to capture the ground truth for outdoor experiments, for instance the structured light technique (Scharstein and Szeliski, 2003). Also the reconstruction of other more stable objects like buildings could be attempted. Furthermore, multi-viewed dense matching can be used to improve the self-training. Multiple images can in fact provide denser reconstruction results; meanwhile a consistency check among more than two images is able to further remove outliers which guarantees more reasonable training data. The self-training strategy of MC-CNN provides the possibility of detailed plant reconstruction and avoids the complexity of collecting ground truth especially in extreme situations.

Acknowledgments

The work was funded by the ‘ForDroughtDet’ project (FKZ: 28WB410602). We are indebted to Dr. Thomas Schneider, Emanuel Jachmann, and Christian Kempf from Technical University of Munich for their continuing support to the data collection. We are also grateful to Tobias Koch at Technical University of Munich who provided expertise for acquiring the LiDAR data. We acknowledge Dr. Miguel Figueiredo Vaz Pato from the German Aerospace Center for contributing to English proof reading and many thanks to the editor and the reviewers for their constructive comments. Yuanxin Xia is supported by a DLR-DAAD Research Fellowship (No. 57265855).

References

Ahmed, O.S., S.E. Franklin, and M.A. Wulder, 2014. Integration of lidar and landsat data to estimate forest canopy cover in coastal British Columbia, Photogrammetric Engineering & Remote Sensing, 80(10): 953-961.

Bolles, R.C., H.H. Baker, and D.H. Marimont, 1987. Epipolar-plane image analysis: An approach to determining structure from motion, International Journal of Computer Vision, 1(1): 7-55.

Bromley, J., J.W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, 1993. Signature verification using a siamese time delay neural network, International Journal of Pattern Recognition and Artificial Intelligence, 7(4): 669-688.

d'Angelo, P., 2016. Improving semi-global matching: cost aggregation and confidence measure, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 41(B1): 299-304.

d'Angelo, P., and P. Reinartz, 2011. Semiglobal matching results on the ISPRS stereo matching benchmark, Proceedings of ISPRS Workshop, Hannover, Germany, 38-4(W19): 79-84.

Freeman, M.P., D.A. Stow, and D.A. Roberts, 2016. Object-based image mapping of conifer tree mortality in San Diego county based on multitemporal aerial ortho-imagery, Photogrammetric Engineering & Remote Sensing, 82(7): 571-580.

Gatziolis, D., J.F. Lienard, A. Vogs, and N.S. Strigul, 2015. 3D tree dimensionality assessment using photogrammetry and small unmanned aerial vehicles, Public Library of Science ONE, 10(9): e0137765.

Geiger, A., P. Lenz, C. Stiller, and R. Urtasun, 2013. Vision meets robotics: The KITTI dataset, International Journal of Robotics Research, 32(11): 1231-1237.

Girardeau-Montaut, D., M. Roux, R. Marc, and G. Thibault, 2005. Change detection on points cloud data acquired with a ground laser scanner, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 36(part 3): W19.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2): 328-341.

Hirschmüller, H., 2011. Semi-global matching - motivation, developments and applications, Proceedings of Photogrammetric Week.

Hirschmüller, H., and D. Scharstein, 2009. Evaluation of stereo matching costs on images with radiometric differences, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(9): 1582-1599.

Joung, S., S. Kim, B. Ham, and K. Sohn, 2017. Unsupervised stereo matching using correspondence consistency, IEEE International Conference on Image Processing, pp. 2518-2522.

Kankare, V., M. Holopainen, M. Vastaranta, E. Puttonen, X. Yu, J. Hyyppä, M. Vaaja, H. Hyyppä, and P. Alho, 2013. Individual tree biomass estimation using terrestrial laser scanning, ISPRS Journal of Photogrammetry and Remote Sensing, 75: 64-75.

Knöbelreiter, P., C. Vogel, and T. Pock, 2018. Self-supervised learning for stereo reconstruction on aerial images, IEEE International Geoscience and Remote Sensing Symposium, pp. 4383-4386.

Krizhevsky, A., I. Sutskever, and G.E. Hinton, 2012. Imagenet classification with deep convolutional neural networks, Proceedings of Advances in Neural Information Processing Systems, pp. 1097-1105.

Kuschk, G., P. d'Angelo, R. Qin, D. Poli, P. Reinartz, and D. Cremers, 2014. DSM accuracy evaluation for the ISPRS Commission I image matching benchmark, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 40(1): 195-200.

Lawrence, S., C.L. Giles, A.C. Tsoi, and A.D. Back, 1997. Face recognition: A convolutional neural network approach, IEEE Transactions on Neural Networks, 8(1): 98-113.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11): 2278-2324.

Levin, S.A., 1999. Fragile Dominion: Complexity and the Commons, Perseus Books, Cambridge, Massachusetts.

Luo, W., A.G. Schwing, and R. Urtasun, 2016. Efficient deep learning for stereo matching, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, pp. 5695-5703.

Mei, X., X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, 2011. On building an accurate stereo matching system on graphics hardware, Proceedings of IEEE International Conference on Computer Vision Workshops, pp. 467-474.

Menze, M., and A. Geiger, 2015. Object scene flow for autonomous vehicles, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, pp. 3061-3070.

Okutomi, M., and T. Kanade, 1993. A multiple-baseline stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(4): 353-363.

Qin, R., X. Huang, A. Gruen, and G. Schmitt, 2015. Object-based 3-D building change detection on multitemporal stereo images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(5): 2125-2137.

Rosu, A.M., M. Pierrot-Deseilligny, A. Delorme, R. Binet, and Y. Klinger, 2015. Measurement of ground displacement from optical satellite image correlation using the free open-source software MicMac, ISPRS Journal of Photogrammetry and Remote Sensing, 100: 48-59.

Scharstein, D., H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, 2014. High-resolution stereo datasets with subpixel-accurate ground truth, German Conference on Pattern Recognition, Münster, Germany.

Scharstein, D., and C. Pal, 2007. Learning conditional random fields for stereo, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, Minnesota, USA, pp. 1-8.

Scharstein, D., and R. Szeliski, 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision, 47(1-3): 7-42.

Scharstein, D., and R. Szeliski, 2003. High-accuracy stereo depth maps using structured light, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, 1: 195-202.

Segal, A., D. Haehnel, and S. Thrun, 2009. Generalized-icp, Proceedings of Robotics: Science and Systems.

Strigul, N., 2012. Individual-based models and scaling methods for ecological forestry: implications of tree phenotypic plasticity, Sustainable Forest Management-Current Research, pp. 359-384.

Tao, S., Q. Guo, S. Xu, Y. Su, Y. Li, and F. Wu, 2015. A geometric method for wood-leaf separation using terrestrial and simulated lidar data, Photogrammetric Engineering & Remote Sensing, 81(10): 767-776.

Tian, J., T. Schneider, C. Straub, F. Kugler, and P. Reinartz, 2017. Exploring digital surface models from nine different sensors for forest monitoring and change detection, Remote Sensing, 9(3): 287.

Wu, Z., D. Dye, J. Vogel, and B. Middleton, 2016. Estimating forest and woodland aboveground biomass using active and passive remote sensing, Photogrammetric Engineering & Remote Sensing, 82(4): 271-281.

Zbontar, J., and Y. LeCun, 2016. Stereo matching by training a convolutional neural network to compare image patches, Journal of Machine Learning Research, 17: 1-32.

Zhou, C., H. Zhang, X. Shen, and J. Jia, 2017. Unsupervised learning of stereo matching, Proceedings of IEEE International Conference on Computer Vision, 2(8): 1567-1575.