The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2022
XXIV ISPRS Congress (2022 edition), 6–11 June 2022, Nice, France

MONOCULAR DEPTH ESTIMATION IN FOREST ENVIRONMENTS

H. Hristova[1,*], M. Abegg[1], C. Fischer[1], N. Rehush[1]

[1] Swiss National Forest Inventory, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Switzerland
(hristina.hristova, meinrad.abegg, christoph.fischer, nataliia.rehush)@wsl.ch

KEY WORDS: Monocular depth estimation, Forestry, Terrestrial laser scanning, Deep learning.

ABSTRACT:

Depth estimation from a single image is a challenging task, especially inside the highly structured forest environment. In this paper, we propose a supervised deep learning model for monocular depth estimation based on forest imagery. We train our model on a new data set of forest RGB-D images that we collected using a terrestrial laser scanner. Alongside the input RGB image, our model uses a sparse depth channel as input to recover the dense depth information. The prediction accuracy of our model is significantly higher than that of state-of-the-art methods when applied in the context of forest depth estimation. Our model brings the RMSE down to 2.1 m, compared to 4 m and above for reference methods.

Figure 1. Depth prediction based on monocular forest images: a) and b) are input and ground truth; c), e) and f) are depth predictions with different methods; d) is absolute error between depth predicted with our model and the ground truth.

1. INTRODUCTION

Knowledge about forest stand characteristics (i.e., the spatial distribution of trees, tree size distribution, etc.) is crucial for forest management and monitoring and assessing the protective function of forests. Besides traditional field sampling used by national forest inventories, other approaches also aid in deriving spatial forest characteristics. In this sense, close-range remote sensing techniques, such as Terrestrial Laser Scanning (TLS), terrestrial photogrammetry, and imaging, have been intensively investigated (Iglhaut et al., 2019).

Traditionally, 3D scene reconstruction is performed using stereo image pairs through triangulation (Ginzler and Hobi, 2015). In recent years, the interest in 3D reconstruction from monocular images (depth estimation from a single image) has increased thanks to the introduction of deep networks. Monocular depth recovery benefits from the newest Deep Learning (DL) architectures and provides an alternative to traditional stereo approaches. Despite this, the success of existing methods and their generalization ability rely on the amount and labeling quality of the training data. Related work on monocular depth estimation trains models on large data sets (Geiger et al., 2013), and achieves outstanding precision on the corresponding test sets. However, the used training sets contain only a limited number of forest images.

In our work, we aim to evaluate the potential of a DL approach for generating depth maps from monocular images collected in a forest environment to later recover the absolute distance from the camera to detectable trees. We show that existing monocular depth estimation approaches fail to produce plausible results for forest imagery due to the highly structured forest environment. By considering the specifics of this environment, we propose a novel framework for depth prediction in forest images. Our method recovers the forest 3D structure from an input RGB image and sparse depth samples. We make use of high-resolution TLS to collect the training data for our model and compute ground-truth depths.

We summarize our main contributions as follows:

- A supervised DL model for monocular depth recovery in forest environments trained on a dedicated set of images.
- A new data set of RGB-D forest images collected using terrestrial laser scanner with an integrated camera.
- An extensive performance evaluation showcasing the advantages of our method over state-of-the-art approaches. Figure 1 hints about the potential of our model.

* Corresponding author

The paper is organized as follows. First, we discuss related work regarding monocular depth estimation. Next, we introduce our new data set of forest imagery and present our model for depth prediction, followed by results and evaluation. Finally, we conclude the paper by discussing future work.

2. RELATED WORK

Research work on monocular depth estimation based on deep learning can be divided into two categories: self-supervised methods leveraging features in the image space to produce relative depth maps and supervised methods that require ground-truth depth and occasionally rely on sparse depth as input (Ming et al., 2021).

RGB-based depth prediction. The first category comprises methods that exploit geometric constraints between stereo pairs or monocular video sequences at training time and use single images to test the model. Garg et al. (Garg et al., 2016) train a Convolution Neural Network (CNN) model to predict depth in an unsupervised manner based on the reconstruction loss between one of the images in the stereo pair and its warped counterpart. This concept is later utilized in (Godard et al., 2017) to learn pixel correspondences between rectified stereo pairs to enable a simultaneous reconstruction of the left image in the pair given the right one and vice versa. The left-right disparity significantly constrains the model trained on Residual Network (ResNet) (He et al., 2016). The predicted relative depth translates into absolute depth with the knowledge of the camera intrinsics. Other methods, such as (Godard et al., 2019), (Zhou et al., 2017), and (Ummenhofer et al., 2017), train jointly a depth network and a pose network. Unlike the approach proposed in (Godard et al., 2017), the camera pose is now unknown. The depth estimation relies on predicting the camera pose from a video sequence by minimizing a reconstruction loss involving consequent video frames. Unsupervised depth estimation requires no ground-truth depths for the input RGB images, offering an advantage over supervised depth estimation. At the same time, the precision of the unsupervised approach is usually lower than the precision of the supervised methods.

RGB-D based prediction and sparse depth. The requirement for additional data in the form of an extra depth channel plays a principal role in increasing the prediction accuracy. The authors in (Eigen et al., 2014) train a multi-scale CNN to first estimate a coarse depth and then refine it locally. A scale-invariant loss function constrains the training by comparing the scale of the ground-truth depth to the prediction. Other supervised methods exploit image features in the RGB-D space, where sparse depth channel D is provided as input to help guide the prediction towards dense depth (Kuznietsov et al., 2017), (Ma and Karaman, 2018). Ma et al. (Ma and Karaman, 2018) include sparse depth in the form of uniformly distributed samples extracted from the ground-truth depth. It is shown that the addition of sparse depth as input improves the model accuracy (Ming et al., 2021).

Datasets. Most related works on depth estimation train models on large data sets, such as KITTI (Geiger et al., 2013) and NYU-Depth V2 (Nathan Silberman and Fergus, 2012), and achieve outstanding precision on the corresponding test sets. The KITTI data set aims to represent broader real-world data. It contains a large number of outdoor stereo video sequences, captured from a moving car using two cameras, and their sparse depth maps. In contrast, NYU-depth-V2 consists of indoor video sequences with measured ground-truth depths. Both data sets contain a limited number of forest images and fail to represent well forest environments. Therefore, depth prediction methods trained on KITTI and NYU-Depth V2 struggle to estimate the depth of the highly structured forest environment.

3. OUR METHOD

In this section, we introduce our supervised method for single-image depth estimation in forest environments. We present our data set of forest imagery that we use to train and test our model, followed by a detailed description of our DL model.

3.1 Data set

Hereafter, we discuss the procedure of creating our data set, from data collection and ground-truth depth post-processing, to data augmentation.

3.1.1 Data collection We created a new data set consisting solely of forest images. To obtain the initial high-quality 360° RGB-D images, we collected data using a FARO Focus 3D 120S terrestrial laser scanner with an integrated camera. We captured the RGB data using the integrated camera and derived the depth channel D from the point cloud data resulting from the scanning. The data were collected from 20 different locations around Switzerland, each of which comprised from 3 to 6 different scanner positions. We then de-noised the point cloud data and extracted the depth information that we merged with the RGB images from the integrated camera. That way, we computed the initial RGB-D 360° images.

3.1.2 Ground-truth depths The scanner captures the RGB information after collecting the point clouds. The latter results in a time lag between the RGB images and the scanned data and may cause inaccurate measurements for non-still objects (*e.g.*, thin branches, leaves).

To compute reliable ground-truth depths, we filtered out the imprecise measurements in the point clouds. To this end, we extracted the VERTICALITY feature using the Cloud Compare software (CloudCompare, 2020). We first computed the VERTICALITY feature with a neighborhood radius of 0.1 m and extracted all points with values above a threshold $t_1 = 0.75$. Next, we re-computed the VERTICALITY feature with a neighborhood radius of 0.5 m on the resulting point cloud and extracted all points above a threshold $t_2 = 0.65$. This series of operations removed most tree crown points from the point cloud, leaving mainly tree stems. To also include ground points, we first extracted the ground by applying the Cloth Simulation Filter (Zhang et al., 2016) on the initial point cloud using the flat scene option and a cloth resolution of 0.2 m. We then merged the tree stem points with the ground points. A 360° RGB visualization of the filtered point cloud data is shown in Figure 2.

We extracted the depth from the filtered point cloud and used it to compute 2D depth maps. To further minimize the amount of remaining noise and outlier pixels, we applied the Median filter (Huang et al., 1979) on the depth maps. For this final de-noising, we tested also a superpixel segmentation (Achanta et al., 2012). The impact of these filtering strategies on the performance of our model is discussed in Section 4.

3.1.3 Data augmentation For computational reasons, we scaled down the initial high-resolution 360° images. For the sake of data augmentation, we rotated the 360° images ten

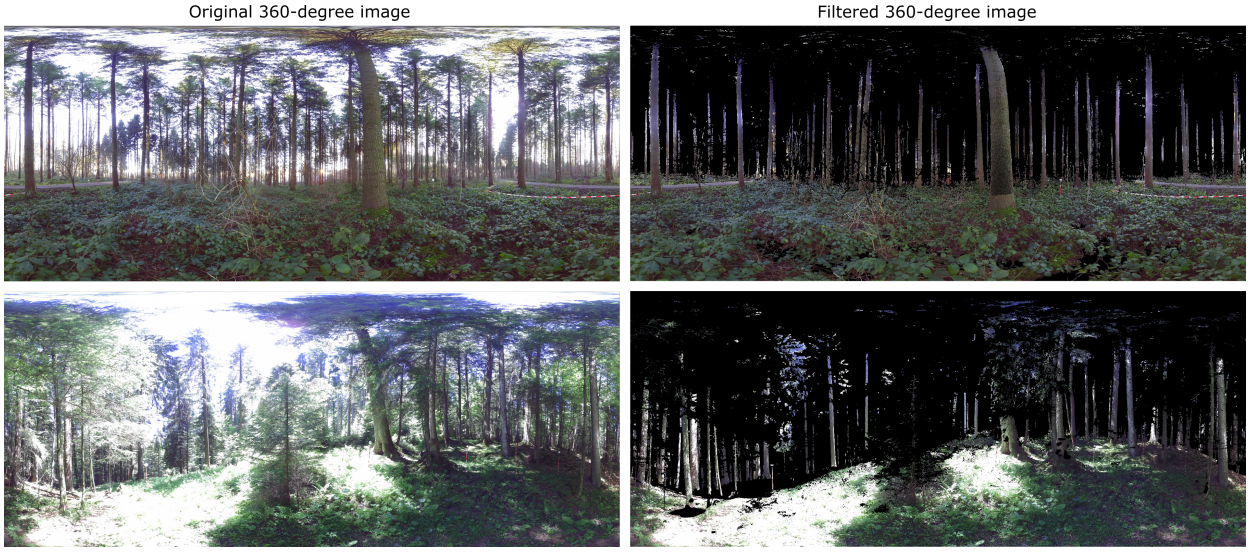

Figure 2. RGB point cloud visualizations of the original TLS data and the de-noised point cloud which was further used as ground truth. As illustrated, the filtered data contains mainly tree stem points and ground points. The second example is more challenging due to the abundant vegetation occluding a major part of the tree stems. Hence the noticeable tree crown areas present in the filtered image.

times by a small angle along the longitude. The latter produced new 360° images as rotated versions of the initial 360° images. Each resulting image was then divided into two parts, overlapping by 50%. Each of them was further cut into two images, resulting in 8760 images with a resolution of 600x700 pixels apiece. Exactly 7120 of these images formed our training data set, whereas 1040 images remained for validation and 600 for testing. The images in our training, validation, and test sets come from different scanning locations. Since we focus on predicting the depth in forest environments and not designing a universal depth estimation method, our model does not require a huge data set. The number of images in the data set is in line with the number of features in our network. Examples of images from our data set are shown in the results in Section 4.

3.2 DL Model

We propose to train a model for predicting depth in monocular images in a supervised manner. To this end, we require the knowledge of reliable and precise ground-truth depth for each input image. The key components of our method are two-fold: uniform depth sampling as input and ground-truth depth constraints in the form of loss function.

3.2.1 Depth sampling The input of our model are RGB-D images. The depth channel includes uniformly distributed random samples from the ground-truth data. Similar to (Ma and Karaman, 2018), the random sampling aims to aid the performance of the model by introducing different number of inputs. Most importantly, feeding sparse depth to the network alongside RGB information reliably guides the depth reconstruction. The input sparse depth contains less than 1% of the pixels in the image. In Section 4, we show the impact of the number of input samples on the prediction accuracy of our model.

3.2.2 Loss function The training is constrained by a loss function consisting of three terms: depth distance L_1, gradient loss L_{grad}, and normal loss L_{norm}.

$$L_1(d, \hat{d}) = \frac{1}{n}\sum_{i}^{n} |d_i - \hat{d}_i| \tag{1}$$

$$L_{grad}(d, \hat{d}) = \frac{1}{n}\sum_{i}^{n} |g_x(d_i) - g_x(\hat{d}_i)| + |g_y(d_i) - g_y(\hat{d}_i)| \tag{2}$$

$$L_{norm}(d, \hat{d}) = 1 - \frac{1}{n}\sum_{i}^{n} \frac{g_x(d_i)g_y(\hat{d}_i)}{\|g(d_i)\|\|g(\hat{d}_i)\|}, \tag{3}$$

where d and $\hat{d}$ are respectively the predicted depth and the ground-truth depth, and $g(g_x(\cdot), g_y(\cdot))$ is the gradient function vector. The gradient and the normal terms are added to the loss function after the fifth epoch.

3.2.3 Network architecture Similar to (Ma and Karaman, 2018), the network architecture that we use to train our model consists of ResNet18 (He et al., 2016) for the feature extraction, followed by a convolution layer with a kernel size of 3x3, and decoding layers. The decoding structure comprises 4 deconvolution layers with a kernel size of 3x3 and a bilinear upsampling layer.

3.2.4 Implementation details The network is implemented in PyTorch and is based on Ma et al.'s software (Ma and Karaman, 2018). The model was trained for 120 epochs with an adaptive learning rate that decreased with 10^{-1} every 20 epochs, starting from an initial value of 0.1. We used a batch size of 15. Color normalization and batch normalization were carried out. We also performed an online data augmentation consisting of random transformations, such as color jitter, image rotation and flipping, and depth scaling, to help the model learn fast by feeding it various image inputs. The training took 9 hours on an NVIDIA TITAN V GPU with 32 GB of RAM.

3.2.5 Evaluation metrics To evaluate the performance of the models, we use three metrics. The standard Root-Mean-Square Error (RMSE) measures the prediction error of the

Figure 3. Predicted depth maps computed using a) our method, b) (Godard et al., 2019), and c) (Ma and Karaman, 2018). Purple areas in all depth maps signify pixels (trees) close to the camera, whereas yellow areas illustrate more distant pixels (trees). Black corresponds to zero-value pixels.

model, whereas δ refers to the threshold accuracy. The metric δ measures the percentage of the pixels whose deviation from the ground truth is less than 1.25 times. Since RMSE may be sensitive to outliers, we also adopt the Mean Absolute Error (MAE) metric to assess the accuracy.

3.2.6 State-of-the-art methods We compared our method against two state-of-the-art methods for monocular depth estimation. As reference methods we chose one RGB-based method (Godard et al., 2019) and one method relying on sparse depth as input (Ma and Karaman, 2018). Ma et al. (Ma and Karaman, 2018) provide two pre-trained models, one with 200 samples trained on ResNet50, and one with more than 13k samples trained on ResNet18. We tested both models and present results from the former one as it showed a better performance on our test set.

4. RESULTS AND DISCUSSION

In this section, we analyse the performance of our model for various input parameters. We also present a visual and metrical comparison between our method and the state-of-the-art approaches. Finally, we discuss limitations.

Performance assessment. The evaluation metrics for our method and the two reference methods, computed on our test set, are shown in Table 1. We analyze the performance of our method for a different number of input depth samples, e.g., 100, 500, and 1000. The RMSE of our model is much lower in comparison to the two reference models, being 2.1 m for our test data set with 500 input samples. We make a similar observation regarding the MAE metric, which is less than 1 m for our model with 500 input samples. Furthermore, the threshold accuracy δ in percentage for our model with 500 input samples is 70%.

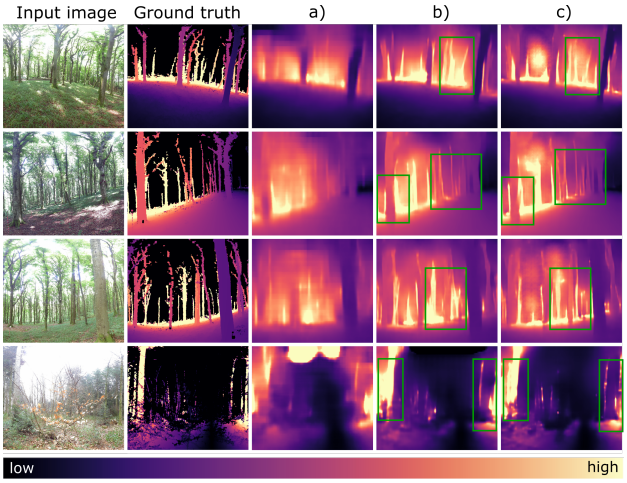

Figure 4. Predicted depth maps computed using our method with a) 100, b) 500, and c) 1000 input depth samples. The differences between a) and b) are easily noticeable. The green boxes show the more subtle differences between b) and c).

We visually compare the outcome of our model for a different number of input depth samples. Figure 4 shows several examples of depth maps predicted using 100, 500, and 1000 input samples. As expected, the higher the number of input samples, the better the visual resemblance to the ground truth. Nonetheless, even with 100 samples, we achieve a threshold accuracy of 64%, a mean error of 2.34 m, and an absolute error of 1.14 m. The best metrics are obtained for 1000 input samples (see Table 1) which is less than 1% of the ground-truth points: RMSE=1.97 m, MAE=0.89 m, δ=72%.

Comparison with state-of-the-art. The reference methods fail to produce plausible results, both numerically and visually. The RMSE of the methods in (Ma and Karaman, 2018) and (Godard et al., 2019) is respectively around 4 m and 4.3 m (see Table 1). A visual comparison of the depth maps, computed with our method and the two state-of-the-art methods for our test set of forest images, is shown in Figure 3. The method in (Ma and Karaman, 2018) has a good threshold accuracy of 49% but fails visually to represent the ground truth. In contrast, the model in (Godard et al., 2019) performs visually well but does not manage to correctly predict the metric depth, with δ of only 15% and a significantly high MAE of 3.6 m.

Our depth maps are visually the closest to the ground truth, exhibiting more errors at tree contours and some distant regions. As observed in the error visualization in Figure 7, our model yields accurate depths for most image parts, including tree stems, especially for ones close to the camera and for tree stems with medium and big diameters. The absolute error for such close objects is below 2 m and increases with the distance (the purple and black areas in Figure 7).

The Figure 5 confirms that, in general, the predicted values are

Method	Input samples	**Less is better** RMSE [m]	**Less is better** MAE [m]	**Higher is better** $\delta < 1.25$
(Godard et al., 2019)	-	4.33	3.58	0.15
(Ma and Karaman, 2018)	200	3.94	1.72	0.49
Our method	100	2.34	1.14	0.64
Our method	500	2.1	0.97	0.7
Our method	1000	1.97	0.89	0.72

Table 1. Evaluation performed on our test set of forest images. The metric δ represents the threshold accuracy of the model. Our method significantly outperforms both state-of-the art methods in terms of RMSE, MAE, and δ for forest environments.

in high agreement with the ground truth, especially within up to 30 m from the camera (see the high point density along the 1:1 line). However, our model also tends to underestimate the depth for some distant objects and overestimate it for some close objects (see also the distributions in the last column in Figure 7).

Both reference methods have difficulties to account for the highly detailed nature of the ground truth, with the method in (Godard et al., 2019) being visually more reliable. For the sake of fair comparison, we note that the two reference methods are trained on data sets that contain a few forest images. Our experiments clearly indicate the need for a model designed specifically for forest environments. Moreover, they illustrates the challenging nature of depth estimation from monocular images collected in such conditions.

Impact of filtering strategies. Our experiments have also shown that the bigger the homogeneous areas in the ground-truth depth, the higher the threshold accuracy δ. However, homogeneity in forest images cannot be guaranteed due to the highly structured nature of the forests. One way to increase homogeneity is through stronger filtering of the ground-truth values. Apart from the Median filter, we tested another filtering strategy: a superpixel segmentation (Achanta et al., 2012) with 1000 segments. Table 2 shows the performance metrics for our method with 500 input samples when different filters were used. The superpixel segmentation outperforms the Median filter in terms of threshold accuracy. The latter comes at the price of higher RMSE and MAE. When no filter is applied, δ drops to 66%, MAE decreases to 0.89 m, and RMSE increases to 2.89 m, indicating the presence of significant outliers.

Limitations. Figure 6 shows two challenging for our model test cases. For both of them, the RMSE is above 3 m. Most of the errors exceeding 3 m come from areas containing leaves and parts of tree crowns. The values from the TLS may be erroneous for such non-static, unstable objects. Despite our de-noising efforts, not all tree crowns and tree leaves were removed, as seen from the second example in Figure 2, causing an increase in the unreliable data and the overall error. The joint distribution of the predicted depth and the ground-truth depth (last column in Figure 6) indicates an over-estimation of the depth for points close to the camera. These points belong to the remaining unfiltered tree leaves, obstructing the tree stems. Yet,

Filter	**Less is better** RMSE [m]	**Less is better** MAE [m]	**Higher is better** $\delta < 1.25$
Median filter	2.1	0.97	0.7
Superpixels	2.35	1.08	0.73
No filter	2.89	0.86	0.66

Table 2. Performance evaluation of our method for different filtering strategies. The metrics were computed on our test set.

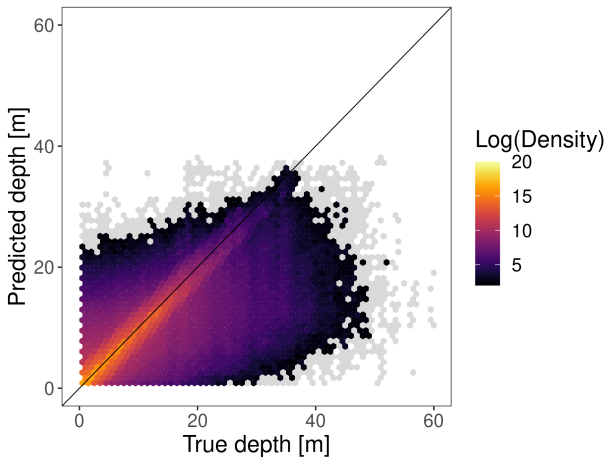

Figure 5. Ground-truth depth versus predicted depth computed on all images from our test set. The yellow and orange areas indicate points with *high* density, whereas the purple and black areas show points with *low* density. The gray areas represent areas with *very low* density (a few depth points). The most densely populated areas lie along the 1:1 line. Often, our model tends to underestimate the depth of distant objects.

the points with the highest density lie along the 1:1 line. As seen from the absolute error maps in the fourth column in Figure 6, our depth prediction remains accurate for most tree stems and ground points, and tree stems precisely are of prior interest to us. However, overcoming the limitation caused by incomplete de-noising would reduce the overall error and help our model learn depth more efficiently.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a supervised deep-learning model for recovering the depth of forest images. We used RGB information and sparse depth samples from a single image to reconstruct its dense 3D structure. We trained the proposed model on our dedicated set of RGB-D forest images. We illustrated that monocular depth estimation inside the forest environment is challenging for reference methods. The latter is partly due to the limited number of forest image examples in existing image data sets. Our performance evaluation has shown that our model outperforms, visually and metrically, state-of-the-art approaches when applied to our test images.

Despite the big potential of monocular depth estimation for RGB-D forest imagery shown in this paper, there is room for improvement. Our model may exhibit high errors at areas away from the camera with insufficient pixel information. The surrounding information may prevent the network from recognizing and extracting meaningful features in such areas. Fu-

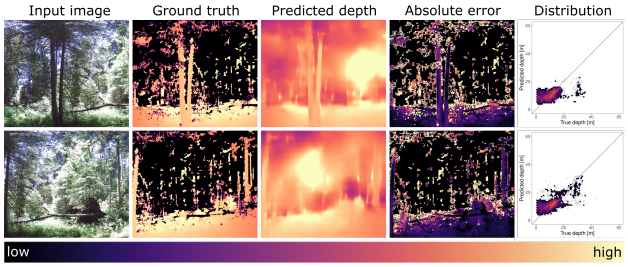

Figure 6. Challenging cases. The predicted depth maps are computed with our method using 500 input depth samples. The absolute errors range in [0m, 3m]. Most yellow areas in the absolute error maps correspond to leaves that were not filtered out during our de-noising procedure. Such areas are overestimated by our model, as shown in the plots in the last column. Yet, the prediction accuracy for tree stems and ground points remains high.

ture work would involve tackling this open question by using deeper networks and introducing appropriate resolution increase strategies to better the feature extraction.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11), 2274–2282.

CloudCompare, Z., 2020. CloudCompare: 3D Point Cloud and Mesh Processing Software.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Garg, R., Kumar, B. V., Carneiro, G., Reid, I., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. *European Conference on Computer Vision*, Springer, 740–756.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*.

Ginzler, C., Hobi, M., 2015. Countrywide stereo-image matching for updating digital surface models in the framework of the swiss national forest inventory, remote sens., 7, 4343–4370.

Godard, C., Mac Aodha, O., Brostow, G. J., 2017. Unsupervised monocular depth estimation with left-right consistency. *CVPR*.

Godard, C., Mac Aodha, O., Firman, M., Brostow, G. J., 2019. Digging into Self-Supervised Monocular Depth Prediction.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, T., Yang, G., Tang, G., 1979. A fast two-dimensional median filtering algorithm. *IEEE transactions on acoustics, speech, and signal processing*, 27(1), 13–18.

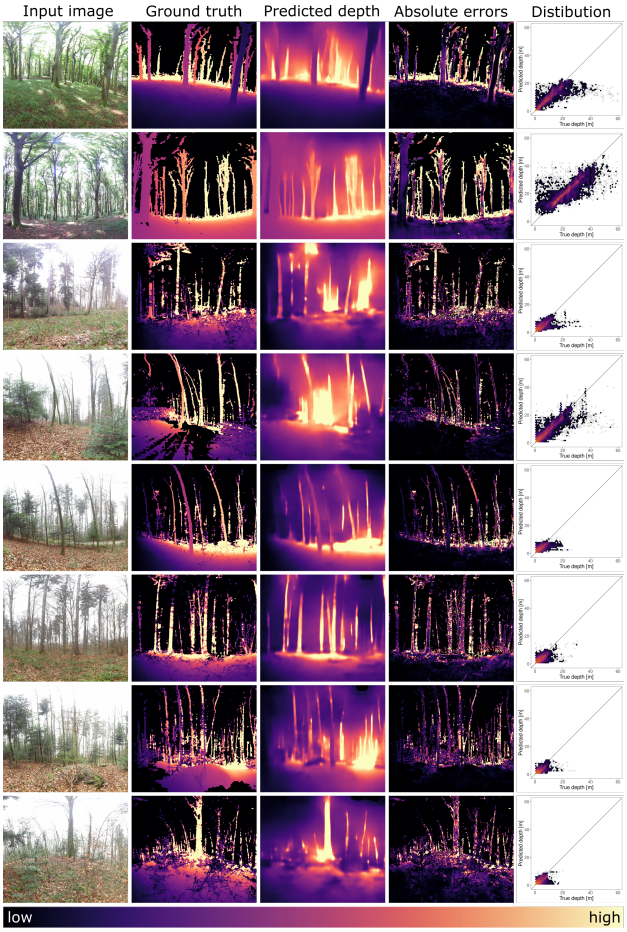

Figure 7. Absolute error (in m) between our predicted depth and the ground-truth depth. Purple areas in columns two and three signify image parts (including trees) close to the camera, whereas yellow areas illustrate more distant image objects. For the fourth column, yellow areas represent absolute errors close to 3 m, whereas black and purple areas signify small absolute errors close to 0 m. The last column shows the joint distribution of predicted depth and ground-truth depth, with yellow and orange points indicating the highest density.

Iglhaut, J., Cabo, C., Puliti, S., Piermattei, L., O'Connor, J., Rosette, J., 2019. Structure from motion photogrammetry in forestry: A review. *Current Forestry Reports*, 5(3), 155–168.

Kuznietsov, Y., Stuckler, J., Leibe, B., 2017. Semi-supervised deep learning for monocular depth map prediction. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6647–6655.

Ma, F., Karaman, S., 2018. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image.

Ming, Y., Meng, X., Fan, C., Yu, H., 2021. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438, 14-33.

Nathan Silberman, Derek Hoiem, P. K., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. *ECCV*.

Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T., 2017. Demon: Depth and motion network

for learning monocular stereo. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5038–5047.

Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X., Yan, G., 2016. An easy-to-use airborne LiDAR data filtering method based on cloth simulation. *Remote sensing*, 8(6), 501.

Zhou, T., Brown, M., Snavely, N., Lowe, D. G., 2017. Unsupervised learning of depth and ego-motion from video. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1851–1858.