Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Semi-Supervised Monocular Depth Estimation with Left-Right Consistency Using Deep Neural Network

Ali Jahani Amiri[1], Shing Yan Loo[1,2], and Hong Zhang[1]
[1]*Department of Computing Science, University of Alberta, Canada*
[2]*Faculty of Engineering, Universiti Putra Malaysia, Malaysia*

***Abstract*— There has been tremendous research progress in estimating the depth of a scene from a monocular camera image. Existing methods for single-image depth prediction are exclusively based on deep neural networks, and their training can be unsupervised using stereo image pairs, supervised using LiDAR point clouds, or semi-supervised using both stereo and LiDAR. In general, semi-supervised training is preferred as it does not suffer from the weaknesses of either supervised training, resulting from the difference in the cameras and the LiDARs field of view, or unsupervised training, resulting from the poor depth accuracy that can be recovered from a stereo pair. In this paper, we present our research in single-image depth prediction using semi-supervised training that outperforms the state-of-the-art. We achieve this through a loss function that explicitly exploits left-right consistency in a stereo reconstruction, which has not been adopted in previous semi-supervised training. In addition, we describe the correct use of ground truth depth derived from LiDAR that can significantly reduce prediction error. The performance of our depth prediction model is evaluated on popular datasets, and the importance of each aspect of our semi-supervised training approach is demonstrated through experimental results. Our deep neural network model has been made publicly available.[1].**

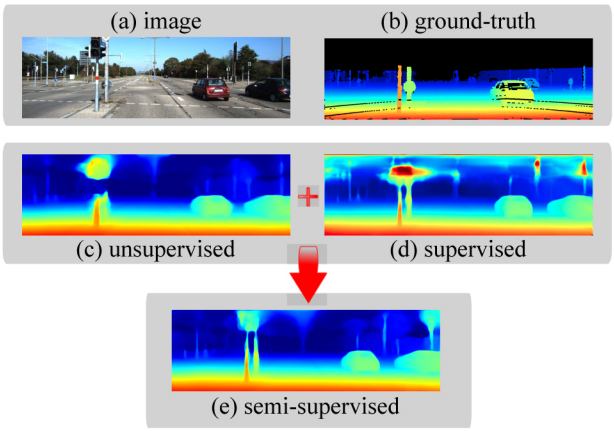


Fig. 1: Using stereo only (c) leads to the noisy depth map. Using LiDAR only (d) results in inaccurate for the top part of the image because there is no ground-truth available. Our semi-supervised method (e) fuses both LiDAR and Stereo and can predict depth more accurately. Ground truth LiDAR (b) has been interpolated for visualization purpose.

I. INTRODUCTION

Single-image depth estimation is an important yet challenging task in the field of robotics and computer vision. A solution to this task can be used in a broad range of applications such as localization of the robot poses [1], [2], 3D reconstruction in simultaneous localization and mapping [3], collision avoidance [4], and grasping [5]. With the rise of deep learning, notable achievements in terms of accuracy and robustness have been obtained in the study of single image depth estimation, and methods of supervised, unsupervised, and semi-supervised have been proposed.

Supervised methods in single-image depth estimation use ground truth derived from LiDAR data. It is time-consuming and expensive to obtain dense ground-truth depth, especially for the outdoor scenes. LiDAR data is also sparse relative to the camera view, and it does not share the same field of view with the camera in general. Consequently, supervised methods are unable to produce meaningful depth estimation in the non-overlapping regions with the image. In contrast, unsupervised methods learn dense depth prediction using the principle of reconstruction from stereo views; hence depth can be estimated for the entire image. However, the accuracy of unsupervised depth estimation is limited by that of stereo reconstruction.

In this paper, we present our research in single-image depth prediction using semi-supervised training that outperforms the state-of-the-art. We propose a novel semi-supervised loss that uses the left-right consistency term originally proposed in [6]. Our network uses LiDAR data for supervised training and rectified stereo images for unsupervised training, and in the testing phase, our network takes only one image to perform depth estimation.

Another focus of our study is the impact of ground truth depth information on the training of our model, when network training is performed with the projected raw LiDAR data and the annotated depth map recently provided by KITTI [7], respectively. We discover that the commonly used projected raw LiDAR contains noisy artifacts due to the displacement between the LiDAR and the camera, leading to poor network performance. In contrast, we use the more reliable preprocessed annotated depth map for training, and we are able to achieve a significant reduction of prediction error.

In summary, we propose in this paper a semi-supervised deep neural network for depth estimation from a single image, with state-of-the-art performance. Our work makes the following three main contributions.

[1]Source code is available at https://github.com/a-jahani/semiDepth

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

- We show the importance of including a left-right consistency term in the loss function for performance optimization in semi-supervised single-image prediction.
- We provide empirical evidence that training with the annotated ground truth derived from LiDAR leads to better depth prediction accuracy than with the raw LiDAR data as ground truth.
- We make our semi-supervised deep neural network - based on the popular Monodepth [6] architecture - available to the community.

The rest of this paper is organized as follows. In Section II, we will review related works to our research, and in Section III we will present our proposed neural network model for single-image depth estimation. Experimental evaluation of our proposed model is provided in Section IV, and conclusion of our work in Section V.

II. Related works

Over the past few years, numerous deep learning based methods have been proposed for the problem of single-image depth estimation. We can divide these deep methods into three categories: supervised, unsupervised, and semi-supervised.

A. Supervised

Supervised methods use ground truth depth, usually from LiDAR in outdoor scenes, for training a network. Eigen et.al. [8] was one of the first who used such a method to train a convolutional neural network. First, they generate the coarse prediction and then use another network to refine the coarse output to produce a more accurate depth map. Following [8], several techniques have been proposed to improve the accuracy of convolutional neural networks such as CRFs [9], inverse Huber loss as a more robust loss function [10], joint optimization of surface normal and depth in the loss function [11]–[13], fusion of multiple depths maps using Fourier transform [14], and formulation of depth estimation as a problem of classification [15].

B. Unsupervised

To avoid laborious ground truth depth construction, unsupervised methods based on stereo image pairs have been proposed [16]. Garg et al. [17] demonstrated an unsupervised method in which the network is trained to minimize the stereo reconstruction loss; i.e., the loss is defined such that the reconstructed right image (i.e., obtained by warping the left image using the predicted disparity) matches the right image. Later on, Godard et al. [6] extended the idea by enforcing a left-right consistency that makes the left-view disparity map consistent with the right-view disparity map. The unsupervised training of our model is based on [6]. Given a left view as input, the model in [6] outputs two disparities of the left view and the right view, while we are outputting only one depth map for one input image in the form of inverse depth instead of disparity. As a result, we treat both left and right images equivalently which allows us to eliminate the overhead of the post-processing step in [6]. By making these changes, our unsupervised model outperforms [6] as will be discussed in Section IV.

C. Semi-Supervised

Unlike unsupervised methods, there has not been much work on semi-supervised learning of depth. Luo et al. [18] and Guo et al. [19] proposed a method that consists of multiple sequential unsupervised and supervised training stages; hence their method could be categorized as a semi-supervised method although they did not use LiDAR and stereo images at the same time in training. Closest to our work is Kuznietsov et al. [20] who proposed adding the supervised and unsupervised loss term in the final loss together resulting in using LiDAR and stereo at the same time in training. One of the main differences between [20] and ours is that we have the left-right consistency term first proposed by [6]. Having this term makes the prediction consistent between left and right. Another difference is that their supervised loss term was directly defined on the depth values whereas we defined it on inverse depth instead. As discussed in [20], a loss term on depth values makes the training unstable because of the high gradients in the early stages of the training. To remedy the situation, Kuznietsov et al. proposed to gradually fade in the supervised loss to achieve convergence whereas our method does not have this problem and does not need to fade in supervised or unsupervised loss terms. In Section IV-C, we show qualitatively and quantitatively that we can obtain better accuracy than [20], which is considered the state-of-the-art in semi-supervised single image depth estimation, as the result of the above considerations.

III. Method

Our approach is based on Monodepth proposed by Godard et al. [6]. Their work is unsupervised and only uses rectified stereo images in training. In this paper, we extend their work and add ground-truth depth data as additional supervision training data. To the best of our knowledge, we are the first one to use left-right consistency proposed by Godard [6] in a semi-supervised framework of single image depth estimation. Fig. 2 shows the different loss terms we use in our training phase, to be described in detail in the next section.

A. Loss Terms

Similar to [6], we define L_s for each output scale s. We calculate loss for four scales. Hence the total loss is defined as $L_{total} = \sum_{s=1}^{4} L_s$.

$$L_s = \lambda_1 E_{reconstruction} + \lambda_2 E_{lr} + \lambda_3 E_{supervised} + \lambda_4 E_{smooth} \quad (1)$$

where λ_i are scalars and the E terms are defined below:

1) Unsupervised Loss $E_{reconstruction}$: We use photometric reconstruction loss between left and right image. Similar to other unsupervised methods, we assume photometric constancy between left-right images. Inverse warping has been used to get the estimated left/right image and then the estimated image is compared with its corresponding real

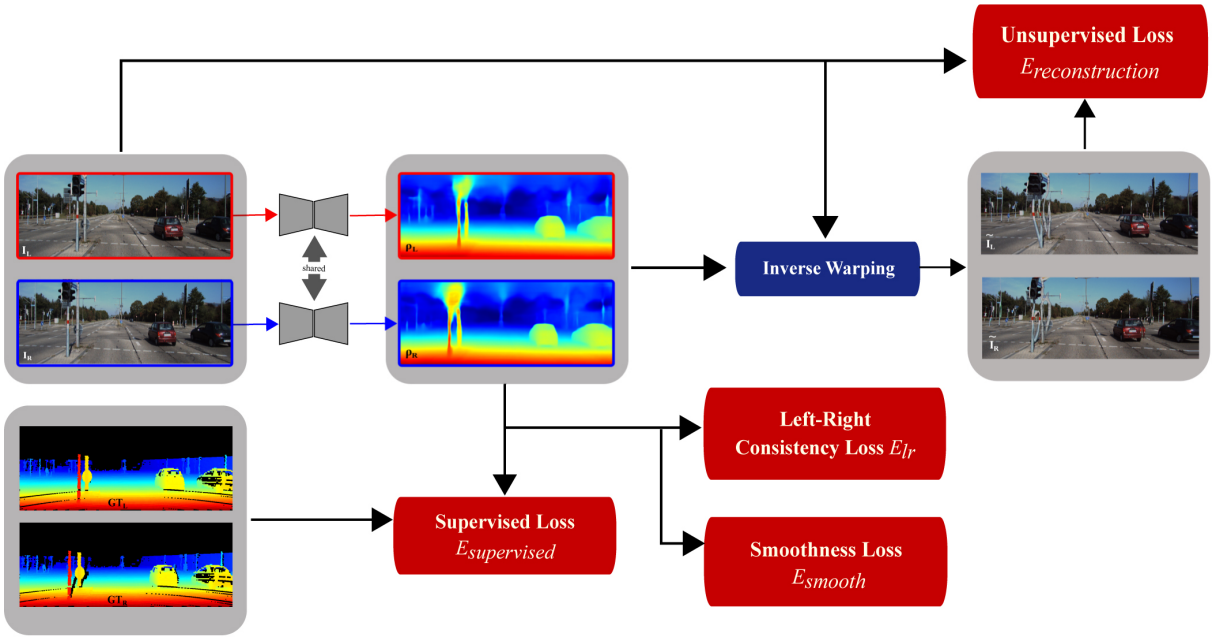


Fig. 2: Overview of the schematic of our proposed loss. There are 4 terms in our loss $E_{reconstruction}$, $E_{supervised}$, E_{lr}, E_{smooth}, $E_{supervised}$. Subscript L and R refers to left and right image, respectively. ρ refers to output of our network inverse depth. We use bilinear sampler in the inverse warping function.

image. In the inverse warping, bilinear sampler is used to make the pipeline differentiable. For comparison, we use the combination of the structural similarity index (SSIM) and L1 used by Godard et al. [6], and the ternary census transform used in [21]–[23]. SSIM and the ternary census transform can compensate for the gamma and illumination change to some extent and result in improved satisfaction of the constancy assumption. Our unsupervised photometric image reconstruction loss term E_u is defined as follows:

$$\begin{aligned} E_{reconstruction} &= \sum_{k\in\{l,r\}} f(I^k, \hat{I}^k) \\ f(I,\hat{I}) &= \frac{1}{N}\sum_{i,j} \alpha_1 * \frac{1-SSIM(I_{ij},\hat{I}_{ij})}{2} + \\ &\quad \alpha_2 * ||I_{ij} - \hat{I}_{ij}||_1 + \\ &\quad \alpha_3 * census(I_{ij}, \hat{I}_{ij}) \end{aligned} \tag{2}$$

where I^l, I^r, $\hat{I}^l$, and $\hat{I}^r$ are the left image, right image and their reconstructed images, respectively. N is the total number of pixels. α_1, α_2, and α_3 are scalars that define the contribution of each term to the total reconstruction loss.

2) left-right depth consistency loss E_{lr}: To ensure the equal contribution of both left and right images in the network training, we feed left and right images independently to the network, and then we jointly optimize the output of the network such that the predicted left and right depth maps are consistent. As explained in [6], left-right depth consistency loss attempts to make the inverse depth of the left (or right) view the same as the projected inverse depth of the right (or left) view. This type of loss is similar to forward-backward consistency for optical flow estimation [21]. We define our left-right depth consistency loss as follow:

$$E_{lr} = \frac{1}{N}\sum_{i,j} ||\rho^l_{ij} - \rho^r_{ij+d^l_{ij}}||_1 + ||\rho^r_{ij} - \rho^l_{ij+d^r_{ij}}||_1, \tag{3}$$

where ρ_l and ρ_r are the predicted inverse depth for left and right images, respectively. d^l and d^r are predicted disparities corresponding to left and right images, respectively. The conversion of inverse depth ρ to disparity d is calculated using (4):

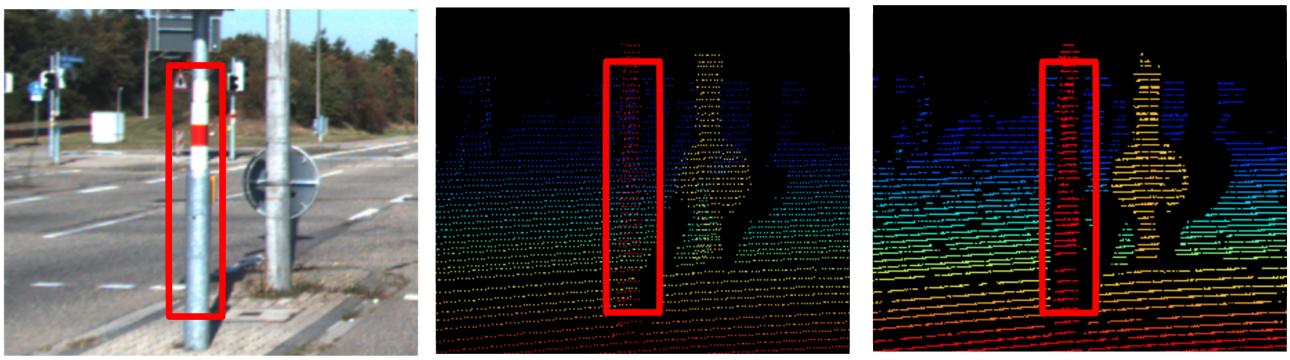
$$d = baseline * f * \rho, \tag{4}$$

where f is the focal length of the camera.

3) Supervised Loss E_s: The supervised loss term measures the difference between the ground truth inverse depth Z^{-1} and the predicted inverse depth ρ for the points Ω where the ground truth is available.

$$E_{supervised} = \sum_{k\in\{l,r\}} \frac{1}{M_k} \sum_{i,j\in\Omega_k} ||\rho^k_{ij} - {Z^{-1}}^k_{ij}||_1 \tag{5}$$

where Ω_l and Ω_r are the points where the ground truth depths are available for the left and right images, respectively. M_l and M_r are the total number of the pixels that ground truth is available for left and right images, respectively.



(a) image (b) projected raw LiDAR (c) annotated depth map

Fig. 3: Qualitative comparison between (b) projected raw LiDAR containing occlusion artifacts due to the displacement between the camera and LiDAR and (c) annotated depth map without any occlusion artifact. We use annotated depth map dataset (c) for our training and evaluation. (b) shows the erroneous depth values for points (green pixels among red for the pole bounded by the red rectangle) that are occluded from the camera point of view but not LiDAR point of view.

4) Smoothness Loss E_{smooth}: As suggested in [6], [20], the smoothness loss term is a regularization term that encourages the inverse depth to be locally smooth with a L_1 penalty on inverse depth gradients. We define our smoothness regularization term as:

$$E_{smooth} = \frac{1}{N} \sum_{k \in \{l,r\}} \sum_{i,j} |\partial_x \rho_{ij}^k| e^{-|\partial_x I_{ij}^k|} + |\partial_y \rho_{ij}^k| e^{-|\partial y I_{ij}^k|} \tag{6}$$

Since the depth is not continuous around object boundaries, this term encourages the neighbouring depth values to be similar in low gradient image regions and dissimilar otherwise.

IV. Experiments

For comparison, we use the popular Eigen split [8] in KITTI dataset [7] that has been used in the previous methods. Using this split, we notice the same problem mentioned by Aleotti et al. [24] that, when LiDAR points are projected into the camera space, an artifact results around objects that are occluded in the image but not from the LiDAR point of view. This is due to the displacement between the LiDAR and the camera sensors. Recently Uhrig et al. [7] provided preprocessed annotated depth maps of KITTI by a preprocessing step on projected raw LiDAR data. They used multiple sequences, left-right consistency checks, and untwisting methods to carefully filter out outliers and densify projected raw LiDAR point clouds. Fig. 3 shows the occlusion artifact in raw projected LiDAR and the corresponding annotated depth map dataset provided by [7]. Since the occlusion artifact is filtered out in the annotated depth ground truth, we train our model with this more accurate ground truth. The first and the third row of Table II show the effect of the training network with the projected raw LiDAR versus the annotated ground truth.

In the rest of the experiments, we evaluate our method based on the official KITTI annotated depth map rather than noisy projected raw LiDAR. Table I contains the quantitative evaluation of the projected raw LiDAR based on the provided annotated depth map ground truth if a depth value of a pixel exists in the both annotated depth map and projected raw LiDAR (54.89% of the LiDAR points,i.e. have been evaluated). The large error for projected raw LiDAR suggests that raw LiDAR is not as accurate as annotated depth maps.

A. Evaluation Metrics

We use the standard metrics used by previous researchers. [6], [8], [20]. Specifically, we use RMSE, RMSE$_{log}$, absolute relative difference (Abs Rel), squared relative difference (Sq Rel), and the percentage of depths (δ) within a certain threshold distance to its ground truth.

B. Implementation Details

We train our network from scratch using Tensorflow [28]. Our network and training procedure are identical to the Resnet50 network used by Godard et al. [6] except for the decoder part in which we have one output instead of two for each scale. As [6] all inputs are resized to 256*512. The output of the network, i.e., inverse depth, is limited to 0 to 1.0 using the sigmoid function. We use Adam optimiser [29] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ with initial learning rate of 10^{-4}, and that remains constant for the first 15 epochs and being halved every 5 epochs for the next 10 epochs for a total of 25 epochs. The hyperparameters for loss are chosen as $\lambda_1 = 1$, $\lambda_2 = 1.0$, $\lambda_3 = 150.0$, $\lambda_4 = 0.1$, $\alpha_1 = 0.85$, $\alpha_2 = 0.15$, and $\alpha_3 = 0.08$.

C. Results

Table I shows the quantitative comparison with the state of the art methods in Eigen split using reliable annotated depth maps for training and testing. Although supervised methods, e.g., DORN [15] can achieve better quantitative performance according to some metrics than semi-supervised methods, they produce an inaccurate prediction of the top portion of the image, which can be seen in Fig. 4, where the LiDAR's field of view is different from that of the camera.

method	type	Dataset	lower is better				higher is better		
			Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
raw LiDAR	-	-	0.010	0.126	1.209	0.054	0.993	0.996	0.998
DORN [15]	S	*K*	**0.080**	**0.332**	**2.888**	**0.120**	**0.938**	**0.986**	**0.995**
SemiDepth(Ours)	S	*K*	0.096	0.552	3.995	0.152	0.892	0.972	0.992
Monodepth [6](Resnet50)	U	*C+K*	0.085	0.584	3.938	0.135	0.916	**0.980**	0.994
MonoGAN [24](Resnet50)	U	*C+K*	0.096	0.699	4.236	0.150	0.899	0.974	0.992
SemiDepth(Ours)	U	*C+K*	**0.082**	**0.551**	**3.837**	**0.134**	**0.920**	**0.980**	**0.993**
Kuznietsov et. al. [20]	Semi	*I+K*	0.089	0.478	3.610	0.138	0.906	0.980	0.995
SVSM FT [18]	Semi	*I+F+K*	**0.077**	**0.392**	3.569	0.127	0.919	0.983	0.995
SemiDepth (full) (Ours)	Semi	*C+K*	0.078	0.417	**3.464**	**0.126**	**0.923**	**0.984**	**0.995**

TABLE I: Quantative evaluation on 652 (93.5%) test images of Eigen Split from the KITTI Dataset. We use official annotated depth map dataset as ground truth instead of noisy projected raw LiDAR. U, S, Semi means unsupervised, supervised and semi-supervised training, respectively. Results of [6], [24] are achieved without the post-processing step. *I*, *C*, *K*, and *F* refer to ImageNet [25], Cityscapes [26], KITTI [7] and FlyingThings3D datasets, respectively. *I* indicates that an encoder is initialized with a pre-trained model trained on ImageNet. All evaluations are using crop from [17]. Depth is capped at 80.0 meters.

Training Data		Loss Term				
Projected Raw LiDAR	Annotated Depth map	Left-Right Consis-tency	Abs Rel	Sq Rel	RMSE	RMSE_{log}
✓		✓	0.120	1.154	5.614	0.204
	✓		0.110	0.973	5.373	0.191
	✓	✓	**0.108**	**0.949**	**5.369**	**0.190**

TABLE II: The effect of the left-right cons:istency term and using annotated depth map in our semi-supuervised training. The results are evaluated on 200 images of KITTI Stereo 2015 split [27]. The second and the third row show that exploiting left-right consistency helps achieving better accuracy. The first and the third row show training on annotated depth map significantly reduces error. The result from our method is shown in bold.

By treating left and right images equivalently and defining our loss symmetrically, we eliminate the post-processing step needed in [6]. As shown in Table I, our unsupervised model outperforms our baseline unsupervised model [6]. In addition, from Table I among the evaluated semi-supervised methods, our method outperforms [20], considered the state-of-the-art, with respect to the majority of the performance metrics. To investigate in detail the effect of using left-right consistency term in the loss function and that of using the annotated LiDAR ground truth, the advantage of our method is confirmed in Table II, where 200 images of KITTI Stereo 2015 split [27] were used in this controlled experiment.

V. CONCLUSION

In this paper, we have presented our approach to semi-supervised training of a deep neural network for single-image depth prediction. Our network uses a novel loss function that uses the left-right consistency term, which has not been used in the semi-supervised training of depth-prediction networks. In addition, we have explained and experimentally confirmed that, for optimal prediction result, in either supervised or semi-supervised training, careful use of the LiDAR data as the ground truth is important. Extensive experiments have been conducted to evaluate our proposed training approach, and we are able to achieve state-of-the-art performance in depth prediction accuracy. Our network model, which is based on Monodepth that is popularly used within the robotics community, is available online for download.

ACKNOWLEDGMENTS

This work has been supported by NSERC Canadian Robotics Network (NCRN). We would like to thank Godard et al. [6] for making their code publicly available.

REFERENCES

[1] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. Cnn-svo: Improving the mapping in semi-direct visual odometry using single-image depth prediction. *arXiv preprint arXiv:1810.01011*, 2018.

[2] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proc. European Conference on Computer Vision (ECCV'18)*, pages 817–833, September 2018.

[3] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *PProc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, pages 6243–6252, 2017.

[4] Punarjay Chakravarty, Klaas Kelchtermans, Tom Roussel, Stijn Wellens, Tinne Tuytelaars, and Luc Van Eycken. Cnn-based single image obstacle avoidance on a quadrotor. In *Proc. IEEE International Conference on Robotics and Automation (ICRA'17)*, pages 6369–6374. IEEE, 2017.

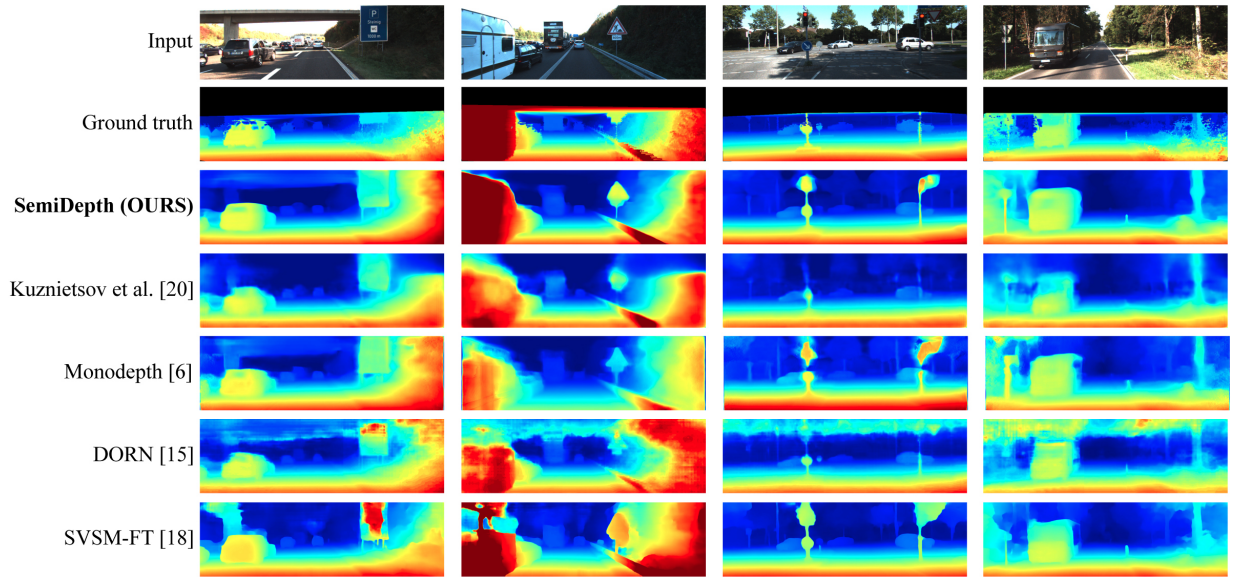


Fig. 4: Qualitative comparison between state-of-the-art methods. We use interpolation in ground truth for visualization purpose.

[5] Deepak Rao, Quoc V Le, Thanathorn Phoka, Morgan Quigley, Attawith Sudsang, and Andrew Y Ng. Grasping novel objects with depth segmentation. In *Proc. IEEE International Conference on Intelligent Robots and Systems (IROS'10)*, pages 2578–2585. IEEE, 2010.

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, pages 270–279, 2017.

[7] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV'17)*, 2017.

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[9] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'15)*, pages 1119–1127, 2015.

[10] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision (3DV'16)*, pages 239–248. IEEE, 2016.

[11] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'15)*, pages 539–547, 2015.

[12] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. *arXiv preprint arXiv:1803.08673*, 2018.

[13] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 283–291, 2018.

[14] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 330–339, 2018.

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 2002–2011, 2018.

[16] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Proc. European Conference on Computer Vision (ECCV'16)*, pages 842–857. Springer, 2016.

[17] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proc. European Conference on Computer Vision (ECCV'16)*, pages 740–756. Springer, 2016.

[18] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 155–163, 2018.

[19] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proc. European Conference on Computer Vision (ECCV'18)*, pages 484–500.

[20] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, pages 6647–6655, 2017.

[21] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. European Conference on Computer Vision (ECCV'94)*, pages 151–158. Springer, 1994.

[23] Fridtjof Stein. Efficient computation of optical flow using the census transform. In *Joint Pattern Recognition Symposium*, pages 79–86. Springer, 2004.

[24] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *15th European Conference on Computer Vision (ECCV) Workshops*, 2018.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[26] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 140:60–76, 2018.

[28] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

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