DA-NET : Monocular Depth Estimation using Disparity maps Awareness NETwork

Antoine Billy, Pascal Desbarats

DA-NET : Monocular Depth Estimation using Disparity maps Awareness NETwork

Antoine Billy [1,2] and Pascal Desbarats[1]
[1]*Laboratoire Bordelais de Recherches en Informatique, Université de Bordeaux, France*
[2]*Innovative Imaging Solutions, Pessac, France*
{*antoine.billy, pascal.desbarats*}*@labri.fr*

Keywords: Monocular Depth Estimation, Disparity Images, Stereo Vision, Convolutional Networks, U-NET.

Abstract: Estimating depth from 2D images has become an active field of study in autonomous driving, scene reconstruction, 3D object recognition, segmentation, and detection. Best performing methods are based on Convolutional Neural Networks, and, as the process of building an appropriate set of data requires a tremendous amount of work, almost all of them rely on the same benchmark to compete between each other : The KITTI benchmark. However, most of them will use the ground truth generated by the LiDAR sensor which generates very sparse depth map with sometimes less than 5% of the image density, ignoring the second image that is given for stereo estimation. Recent approaches have shown that the use of both input images given in most of the depth estimation data set significantly improve the generated results. This paper is in line with this idea, we developed a very simple yet efficient model based on the U-NET architecture that uses both stereo images in the training process. We demonstrate the effectiveness of our approach and show high quality results comparable to state-of-the-art methods on the KITTI benchmark.

1 INTRODUCTION

The problem of estimating depth in an image got addressed by many authors as a crucial factor in scene understanding applications like, for instance, autonomous driving (Urmson et al., 2008), robot assistance (Cunha et al., 2011), object detection (Volkhardt et al., 2013) or augmented reality systems (Ong and Nee, 2013). Being able to generate a dense depth map from a 2D image turned out to be strongly appealing as modern solutions require a specific equipment such as LiDAR (Schwarz, 2010) in which costs might become expensive, RGB-D camera (Henry et al., 2014) only sustainable for indoor scenarios or stereo vision systems (Bertozzi and Broggi, 1998) with a decreasing precision for large-scale scenes. Predicting depth from an image is relatively easy for a human brain thanks to perspective and the known size of ordinary objects but it can still be deceived by very simple optical illusions. Monocular depth estimation by a computer has thus proven to be a key challenge in recent studies and several benchmarks[1] appeared during the last decade, requiring more precision and accuracy for the incoming methods.

[1]http://www.cvlibs.net/datasets/kitti/eval_depth.php

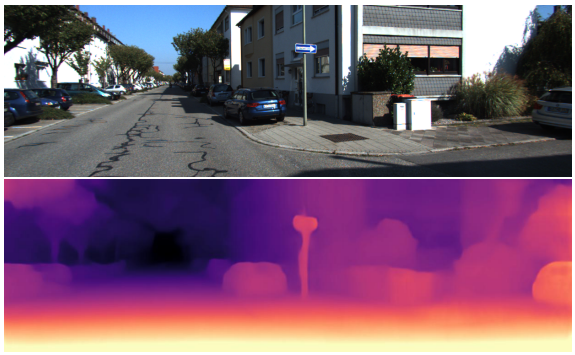


Figure 1: Example output from our model. From Top to Bottom: Original image and its predicted depth.

Thanks to these benchmarks, it is obvious to see that all of the best performing methods (Fu et al., ; Díaz, ; Ren et al., 2019) make use of deep learning approaches, and more specifically Convolutional Neural Networks. To do so, a large amount of data is required for the models to be trained. Even with self-supervised methods (Godard et al., 2017; Zhu et al., 2018; Zhong et al., 2017) the need of either a stereo system or a strict acquisition process with strongly overlapping frames is needed. Building such a database is a colossal job, this is why

most of modern methods use the already existing data sets. This moreover allows these methods to compare themselves with each other on the same scenarios. These data sets provide several sequences of images with their corresponding depth as ground truth. There is a limited number of solutions allowing the data builder to acquire this information. Synthetic dataset (Mayer et al., 2016; Billy et al., 2019) have been developed to ensure perfect ground truth generation. These images however lack of realism and fail to depict real world artifacts regularly occurring in external acquisitions. The use of a RGB-D camera such as the Kinect (Zhang, 2012) has been exploited in (Silberman et al., 2012). This sensor is suitable for close-range indoor scenarios but can not be applied for large outdoor scenes. LiDAR sensors appear to be the best candidates for our application but requires high entry costs. Moreover, a LiDAR sensor will generate a sparse depth map that will need both realignment and interpolation processes. Finally, a stereo system might be embedded to generate dense depth map. This layout is yet not enough on its own because of the lack of precision increasing with the scene depth. (Geiger et al., 2012) took advantage of the last two solutions, combining the strength of the LiDAR and the consistency of stereo vision models, to build a widely used dataset : The so-called KITTI benchmark. Despite the fact that the data from KITTI allow us to use both LiDAR and stereo acquisitions, almost all of the existing methods only use the LiDAR depth map for the training. The role of this paper is to show that by taking into account both stereo images and building their corresponding disparity map, we can perform state of the art results with a very simple network architecture.

2 RELATED WORKS

Depth prediction from single view has gained increasing attention in the computer vision community thanks to the recent advances in deep learning. Simultaneous Localization and Mapping (Thrun, 2008) or Structure from Motion (Ullman, 1979) methods turned out to be powerful tools able to accomplish this mission but will require several images of the same scene with an important overlap between two consecutive frames in order to work efficiently and can not be classified in the single image depth estimation category. Classic depth prediction approaches employ hand-crafted features and probabilistic graphical models (Delage et al., 2007), (Mutimbu and Robles-Kelly, 2013), (Saxena et al., 2007) to yield regularized depth maps, usually making strong assumptions on the scene geometry. Moreover, formulating depth estimation as a Markov Random Field (MRF) learning problem as in (Schwarz, 2010) might result to some issues. As exact MRF learning and inference are intractable in general, most of these approaches employ approximation methods such as multiconditional learning (MCL) or particle belief propagation (PBP). These approximations require complex modeling process and limit their applications to very specific scenarios. Furthermore, recently developed deep convolutional architectures significantly outperformed previous methods in terms of depth estimation accuracy, speed and versatility.

An exhaustive review of CNN single-image depth estimator has been made by (Koch et al.,) in addition with a standardized evaluation protocol. One of the first truly competitive CNN for single-image depth prediction has been developed by (Eigen et al., 2014) using a two scale deep network. Unlike most other previous work in single image depth estimation, they do not rely on hand crafted features or an initial oversegmentation and instead learn a representation directly from the raw pixel values. Several works have built upon the success of this approach using techniques, improving the overall precision of the estimation. These approaches however rely on having high quality, pixel aligned, ground truth depth at training time, and are only based on LiDAR velodyne data. (Zhu et al., 2018; Zhong et al., 2017) proposed self-supervised methods whose training phase does not require ground truth data. Similarly, (Garg et al., 2016) offered an unsupervised CNN for depth prediction. They train a network for monocular depth estimation using an image reconstruction loss. However, their image formation model is not fully differentiable and thus have to perform a Taylor approximation to linearize their loss, resulting in an objective that is more challenging to optimize and again do not take fully advantage of the stereo data.

This principle have been exploited with the build of the DispNet architecture (Mayer et al., 2016), later improved by (Godard et al., 2017). They introduce the integration of disparity in their loss function by taking account of the synthetic depth map developed by (Mayer et al., 2016) and the left-right consistency employed in (Godard et al., 2017). Our approach is based on the use of unsupervised cues such as (Godard et al., 2017) and use both images in the stereo pair equivalently to define our loss function but use the strength of supervised methods to enhance the generated results. A quantitative comparison between our methods and (Godard et al., 2017) is detailed in a further section, showing that we outperform their prediction in almost all metrics.

3 THE DA-NET MODEL

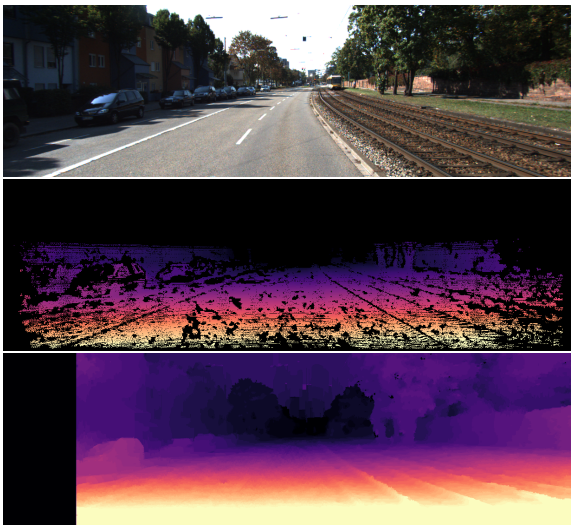


Figure 2: Example data given by KITTI. Top to Bottom: original RGB image, LiDAR depth map, Computed disparity map.

This section describes our single image depth prediction network. We introduce a new depth estimation training loss, including a Disparity Map awareness term. This allow us to train our model with fully dense images. Indeed, as shown in figure 2, the depth image generated by LiDAR acquisition is very sparse and might need strong interpolation to generate a final dense image. This interpolation process generates a rough interpretation of the acquired scene and could easily miss fine details or even small objects such as road signs, trees or,far away, even pedestrians or cars.

On the other hand, a disparity map generated by a stereo vision algorithm will directly generate a depth value for every pixel in the image, resulting in a denser depth estimation. A thorough study of stereo vision algorithms has been exposed by (Scharstein et al., 2001) and several approaches might be used to generate the finest disparity map from rectified images. Semi Global matching (SGBM) (Hirschmüller, 2008) is known to produce more accurate depths and is an integral part of many of the state-of-the-art stereo algorithms. We train a CNN to model the complex non-linear transformation which converts an image to a depth-map. The loss we use for learning this CNN is the photometric difference between the input image, and the inverse warped target image (the other image in the stereo pair). This loss is highly correlated with the prediction error as it can be used to accurately rank two different depth-maps even without using ground-truth labels.

Training Loss Similar to (Godard et al., 2017), we also formulate our problem as the minimization of a photometric reprojection error at training time. Our final loss L (4) is computed as a combination of three main terms.

Photometric Loss Following (Pillai et al., 2019), the similarity between the target image I_t and the synthesized target image image $\hat{I}_t$ is computed using the Structural Similarity (SSIM) (Pillai et al., 2019) term combined with a $L1$ norm as shown in 3:

$$L_p = \alpha_1 \frac{1 - \text{SSIM}(I_t, \hat{I}_t)}{2} + (1 - \alpha_1) \left\| I_t - \hat{I}_t \right\| \quad (1)$$

Smoothness Loss In order to regularize the disparities in textureless low-image gradient regions, we incorporate an edge-aware term based on the generated disparity map. The effect of each of the pyramid-levels is decayed by a factor of 2 on downsampling, starting with a weight of 1 for the 0^{th} pyramid level.

$$L_s = |\delta_x \delta_t| e^{-|\delta_x I_t|} + |\delta_y \delta_t| e^{-|\delta_y I_t|} \quad (2)$$

Disparity Awareness Loss Finally, our main contribution is in the integration of our Disparity Awareness loss. This loss is extremely similar to the first criterion, however, instead of comparing the generated image I_t with the LiDAR output $\hat{I}_t$, we compute its difference with the generated disparity map D_t.

$$L_d = \alpha_2 \frac{1 - \text{SSIM}(I_t, D_t)}{2} + (1 - \alpha_2) \left\| I_t - D_t \right\| \quad (3)$$

Final Loss Ultimately, a binary mask is applied both the the LiDAR and Disparity images. This mask ensures that non relevant pixels (LiDAR wholes or left-most pixels on the disparity map) don't interfere in the final loss computation.

$$L = \alpha L_p + \beta L_d + \gamma L_s \quad (4)$$

Whereas L_p forces the reconstructed image to be similar to the LiDAR input and L_d to the Disparity map, L_s smoothes the generated depth estimation.

3.1 Network Architecture

Our depth estimation network is based on the general U-Net architecture (Ronneberger et al., 2015). As shown in Figure 3, this architecture enables to represent the input image at different scales, while taking advantage of high level features for each step. Although such method has been introduced for medical

image segmentation, we demonstrate hereafter how it can be efficiently used for monocular depth prediction when trained using our proposed loss function.

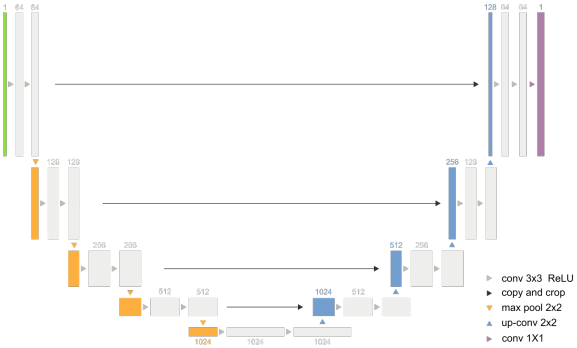


Figure 3: Our network is based on the general U-Net architecture.

4 EXPERIMENTS AND RESULTS

In this section, we compare the performances of our approach to several state-of-the-art methods. We applied straightforward data augmentation by horizontally flipping the images and performed random gamma, brightness, and color shifts by sampling from uniform distributions for each color channel separately. We compare two different variants of our method, with (**ours**) and without (**ours No DA**) Disparity Awareness to evaluate its impact. We evaluate our models, on the Eigen data split KITTI to allow comparison with previously published monocular methods.

4.1 The KITTI dataset

The KITTI dataset by (Geiger et al., 2012) was introduced in IJRR in 2013. Instead of utilizing the entire KITTI dataset, it is common to follow the Eigen split. This split contains 23,488 images from 32 scenes for training and 697 images from 29 scenes for testing. We use the data split of (Eigen et al., 2014). Except in ablation experiments, for training we follow Zhou et al.'s pre-processing to remove static frames. This results in 39,810 monocular images for training and 4,424 for validation. We use the same intrinsic parameters for all images, setting the principal point of the camera to the image center and the focal length to the average of all the focal lengths in KITTI. For fair comparison with state-of-the-art single view depth prediction, we evaluate our results on the same cropped region of interest.

4.2 Evaluation protocol

In addition with the dataset, a development kit is also given by KITTI to evaluate our methods. The following metrics are computed in this kit:

- RMSE: $\sqrt{\frac{1}{T}\sum_{i\varepsilon T}^{T}\left\|d_i - d_i^{gt}\right\|^2}$
- RMSE log: $\sqrt{\frac{1}{T}\sum_{i\varepsilon T}^{T}\left\|\log(d_i) - \log(d_i^{gt})\right\|^2}$
- Sq. relative: $\frac{1}{T}\sum_{i\varepsilon T}^{T}\frac{\left\|d_i - d_i^{gt}\right\|^2}{d_i^{gt}}$
- Abs. relative: $\frac{1}{T}\sum_{i\varepsilon T}^{T}\frac{\left|d_i - d_i^{gt}\right|}{d_i^{gt}}$
- Accuracies: % of d_i s.t. $\max(\frac{d_i}{d_i^{gt}}, \frac{d_i^{gt}}{d_i}) = \delta < \text{thr}$

Even though these statistics are good indicators for the general quality of predicted depth maps, they could be delusive. Particularly, the standard metrics are not able to directly assess the planarity of planar surfaces or the correctness of estimated plane orientations. Furthermore, it is of high relevance that depth discontinuities are precisely located, which is not reflected by the standard metrics. This allows us yet to easily compare our results with others approaches as shown in Tables 1 and 2.

4.3 Results analysis

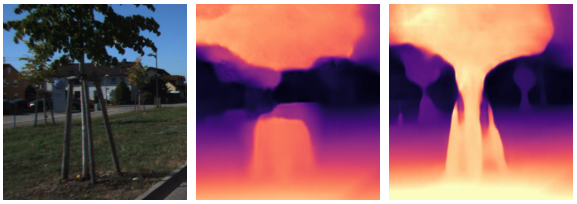
Quantitative results in Table 1 unambiguously highlight the competitiveness of our method with respect to state-of-the-art approaches.

Our scores outperform all the tested methods in every pixel-wise metrics that average the offset between the prediction and the ground truth. This proves that our proposed model provides more accurate depth prediction compared to other methods. This is confirmed by the accuracy evaluation where our method provides better scores for both $\delta = 1.25^2$ and $\delta = 1.25^3$. However, for $\delta < 1.25$ (Godard et al., 2017) produces better results. This is due to the fact that the depth maps generated by SGBM tend to be less precise as the depth increases. Indeed, the depth value of a pixel is directly correlated with the offset from its correspondence in the second image. As this offset is a discrete value in pixel unit, pixels corresponding to faraway objects only have a limited set of values. In contrast, foreground pixels are significantly easier to distinguish, allowing a visible improvement of state-of-the-art results as illustrated in Table 2.

Here, we can see that SGBM tends to create artifacts in the reconstruction, as the consistency of the disparity map lacks of regularization. The method

presented in (Godard et al., 2017) succeeds in generating smoother results. However, fine structures such as tree trunks or sign post might disappear in the reconstruction. Our approach achieves in both smooth prediction with no artifacts, thanks to the combination of photometric loss and disparity awareness, as well as fine structure recovering as the network encoder-decoder architecture is able to handle a large variety of structure scale during prediction.

This scenario is outlined even bolder in Figure 4 where a comparison is made between (Godard et al., 2017) and our method. This points out expressly the advantage of using disparity maps. The tree is clearly visible in the input image but its outlines are partially erased in (Godard et al., 2017) cutting the tree in two halves. Thanks to the Disparity Awareness, our methods successfully segments the contours and the predicted depth fits the real world scenario.



(a) Input image (b) Godard et al. (c) Our method

Figure 4: Qualitative results on KITTI. These zoomed depth maps show that our approach visually outperforms (Godard et al., 2017) for the foreground objects segmentation, producing better results thanks to its Disparity Awareness.

5 CONCLUSION

In this paper we have presented a simple yet efficient deep neural network for monocular depth estimation. We proposed a novel photometric loss that is able to take advantage of disparity map consistency while ensuring the regularity of the predicted depth image. Along with the proposed loss, we showed that the use of both stereo images to generate a dense disparity map instead of just the one given by the LiDAR sensor is a strong tool to predict satisfying dense depth maps that even outperform qualitatively and quantitatively state-of-the-art methods. Further works will include evaluation on other datasets as well as studies on temporal consistency to apply our methods to videos.

REFERENCES

Bertozzi, M. and Broggi, A. (1998). GOLD: A parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE Transactions on Image Processing*, 7(1):62–81.

Billy, A., Pouteau, S., Desbarats, P., Chaumette, S., and Domenger, J. P. (2019). Adaptive slam with synthetic stereo dataset generation for real-time dense 3d reconstruction. In *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 840–848.

Cunha, J., Pedrosa, E., Cruz, C., Neves, A., and Lau, N. (2011). Using a depth camera for indoor robot localization and navigation. In *DETI/IEETA-University of Aveiro*.

Delage, E., Lee, H., and Ng, A. Y. (2007). Automatic single-image 3d reconstructions of indoor Manhattan world scenes. *Springer Tracts in Advanced Robotics*, 28.

Díaz, R. Soft Labels for Ordinal Regression. Technical report.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 3, pages 2366–2374.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. Deep Ordinal Regression Network for Monocular Depth Estimation. Technical report.

Garg, R., Vijay Kumar, B. G., Carneiro, G., and Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9912 LNCS, pages 740–756.

	Lower is better				Higher is better		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train Set Mean	0.361	4.826	8.102	0.377	0.638	0.804	0.894
(Eigen et al., 2014)	0.214	1.605	6.563	0.292	0.673	0.884	0.957
(Yang et al., 2018)	0.198	1.202	5.977	0.266	0.72	0.901	0.932
(Godard et al., 2017)	0.148	1.344	5.927	0.247	**0.803**	0.922	0.964
Ours No DA	0.166	1.401	6.321	0.284	0.701	0.899	0.918
Ours	**0.112**	**0.961**	**5.641**	**0.223**	0.786	**0.944**	**0.971**

Table 1: Quantitative results. Comparison of our method to existing methods on the KITTI Eigen split. Best results are shown in **bold**. Except for (Eigen et al., 2014), all the results have been directly taken from their papers, as we use the exact same split for evaluation. This table shows that our method outperforms state of the art approaches in almost all metrics.

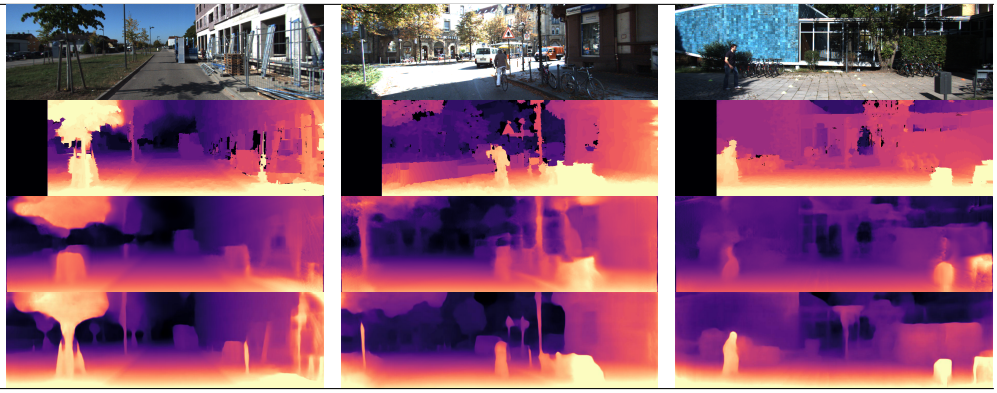
Qualitative results on the KITTI Eigen Split			
Method	Scenario 1	Scenario 2	Scenario 3
Input image			
(Hirschmüller, 2008)			
(Godard et al., 2017)			
Ours			

Table 2: Qualitative results on KITTI. Our method produces superior visual results on distinct scenarios thanks to its Disparity Awareness. Smaller foreground objects such as trees or road signs are clearly better detected in our method.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3354–3361.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6602–6611.

Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. (2014). RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In *Springer Tracts in Advanced Robotics*, volume 79, pages 477–491. Springer Verlag.

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

Koch, T., Liebel, L., Fraundorfer, F., and Körner, M. Evaluation of CNN-based Single-Image Depth Estimation Methods. Technical report.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. (2016). A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 4040–4048.

Mutimbu, L. and Robles-Kelly, A. (2013). A relaxed factorial Markov random field for colour and depth estimation from a single foggy image. In *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings*, pages 355–359.

Ong, S. and Nee, A. (2013). *Virtual and augmented reality applications in manufacturing*.

Pillai, S., Ambruş, R., and Gaidon, A. (2019). SuperDepth: Self-supervised, super-resolved monocular depth estimation. In *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2019-May, pages 9250–9256.

Ren, H., El-khamy, M., and Lee, J. (2019). Deep Robust Single Image Depth Estimation Neural Network Using Scene Understanding.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. Springer Verlag.

Saxena, A., Sun, M., and Ng, A. Y. (2007). Learning 3-D scene structure from a single still image. In *Proceedings of the IEEE International Conference on Computer Vision*.

Scharstein, D., Szeliski, R., and Zabih, R. (2001). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings - IEEE Workshop on Stereo and Multi-Baseline Vision, SMBV 2001*, pages 131–140.

Schwarz, B. (2010). industry perspective technology focus lIDAR Mapping the world in 3D. Technical report.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7576 LNCS, pages 746–760.

Thrun, S. (2008). Simultaneous localization and mapping.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 203(1153):405–426.

Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M. N., Dolan, J., Duggins, D., Galatali, T., Geyer, C., Gittleman, M., Harbaugh, S., Hebert, M., Howard, T. M., Kolski, S., Kelly, A., Likhachev, M., McNaughton, M., Miller, N., Peterson, K., Pilnick, B., Rajkumar, R., Rybski, P., Salesky, B., Seo, Y.-W., Singh, S., Snider, J., Stentz, A., Whittaker,

W. Wolkowicki, Z., Ziglar, J., Bae, H., Brown, T., Demitrish, D., Litkouhi, B., Nickolaou, J., Sadekar, V., Zhang, W., Struble, J., Taylor, M., Darms, M., and Ferguson, D. (2008). Autonomous driving in urban environments. *Journal of Field Robotics*, 25(8):425–466.

Volkhardt, M., Schneemann, F., and Gross, H. M. (2013). Fallen person detection for mobile robots using 3D depth data. In *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pages 3573–3578.

Yang, N., Wang, R., Stuckler, J., and Cremers, D. (2018). Deep Virtual Stereo Odometry : Monocular Direct Sparse Odometry. *European Conference on Computer Vision*, pages 1–17.

Zhang, Z. (2012). Microsoft kinect sensor and its effect.

Zhong, Y., Dai, Y., and Li, H. (2017). Self-Supervised Learning for Stereo Matching with Self-Improving Ability.

Zhu, A., Yuan, L., Chaney, K., and Daniilidis, K. (2018). EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. Robotics: Science and Systems Foundation.