MONOCULAR SEGMENT-WISE DEPTH: MONOCULAR DEPTH ESTIMATION BASED ON A SEMANTIC SEGMENTATION PRIOR

Amir Atapour-Abarghouei[1] *and Toby P. Breckon*[1,2]
Department of {Computer Science[1], Engineering[2]}, Durham University, UK

ABSTRACT

Monocular depth estimation using novel learning-based approaches has recently emerged as a promising potential alternative to more conventional 3D scene capture technologies within real-world scenarios. Many such solutions often depend on large quantities of ground truth depth data, which is rare and often intractable to obtain. Others attempt to estimate disparity as an intermediary step using a secondary supervisory signal, leading to blurring and other undesirable artefacts. In this paper, we propose a monocular depth estimation approach, which employs a jointly-trained pixel-wise semantic understanding step to estimate depth for individually-selected groups of objects (segments) within the scene. The separate depth outputs are efficiently fused to generate the final result. This creates more simplistic learning objectives for the jointly-trained individual networks, leading to more accurate overall depth. Extensive experimentation demonstrates the efficacy of the proposed approach compared to contemporary state-of-the-art techniques within the literature.

***Index Terms*—** Monocular Depth Estimation, Convolutional Neural Networks, Semantic Segmentation

1. INTRODUCTION

As 3D scene understanding is gaining increasing significance within computer vision applications, accurate and efficient depth estimation is now an integral part of many such systems. While strategies such as stereo correspondence [1], structure from motion [2] and depth from shading and light diffusion [3, 4] have produced promising results, prevalent issues such as missing values (holes), depth inhomogeneity and computationally intensive processing or calibration requirements are ubiquitous within such approaches [5]. This has given rise to the necessity of depth refinement post estimation [6, 7, 8]. Recently, novel monocular depth estimation techniques have emerged as a more effective, economical and innovative alternative and have received remarkable attention within the research community [9, 10, 11, 12].

In this work, we propose a model that estimates scene depth based on a single RGB image (Figure 1) by first semantically understanding the scene and then using the knowledge to generate depth for carefully selected segments, i.e. groups of scene objects. These generated segment-wise depth images are subsequently fused by means of a simple summation operation and the overall consistency is controlled by means of an adversarial training procedure as the final step.

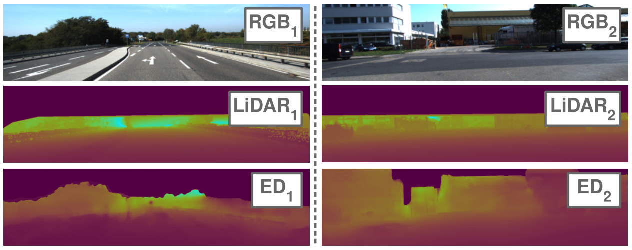

Fig. 1: An example of the results of our monocular depth estimation approach (ED) compared to LiDAR ground truth.

Such a training process is enabled by the use of a large-scale publicly-available dataset of synthetic images [13], which contains pixel-wise ground truth semantic object labels as well as pixel-perfect synthetic depth. Separating the depth estimation process for different object groups during training results in simpler learning objectives for the overall model leading to improved depth estimation accuracy. This results in superior performance compared to some of the most highly-acclaimed approaches within the literature, as demonstrated by our extensive evaluation (Section 4).

The major contributions of this work are thus as follows:

- Using pixel-level scene segmentation as a prior to enhance the performance of monocular depth estimation.
- Utilizing an end-to-end training procedure for an overall model capable of estimating depth for individual groups of scene objects based on a semantic segmentation step jointly trained within the same model.
- A monocular depth estimation approach capable of producing accurate dense scene depth.

2. PRIOR WORK

With the emergence of learning-based approaches, significant improvements have been made to the state of the art in the field of monocular depth estimation in recent years. For instance, in [10], depth is generated by means of a two-scale network trained on RGB and depth. Other approaches [14, 15] have also achieved impressive results using a training procedure directly supervised on real-world depth images despite the scarcity of ground truth depth for supervision.

Recent work has circumnavigated the need for ground truth depth by calculating disparity by reconstructing corresponding views within a stereo framework without ground

Source code will be made available post review.

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

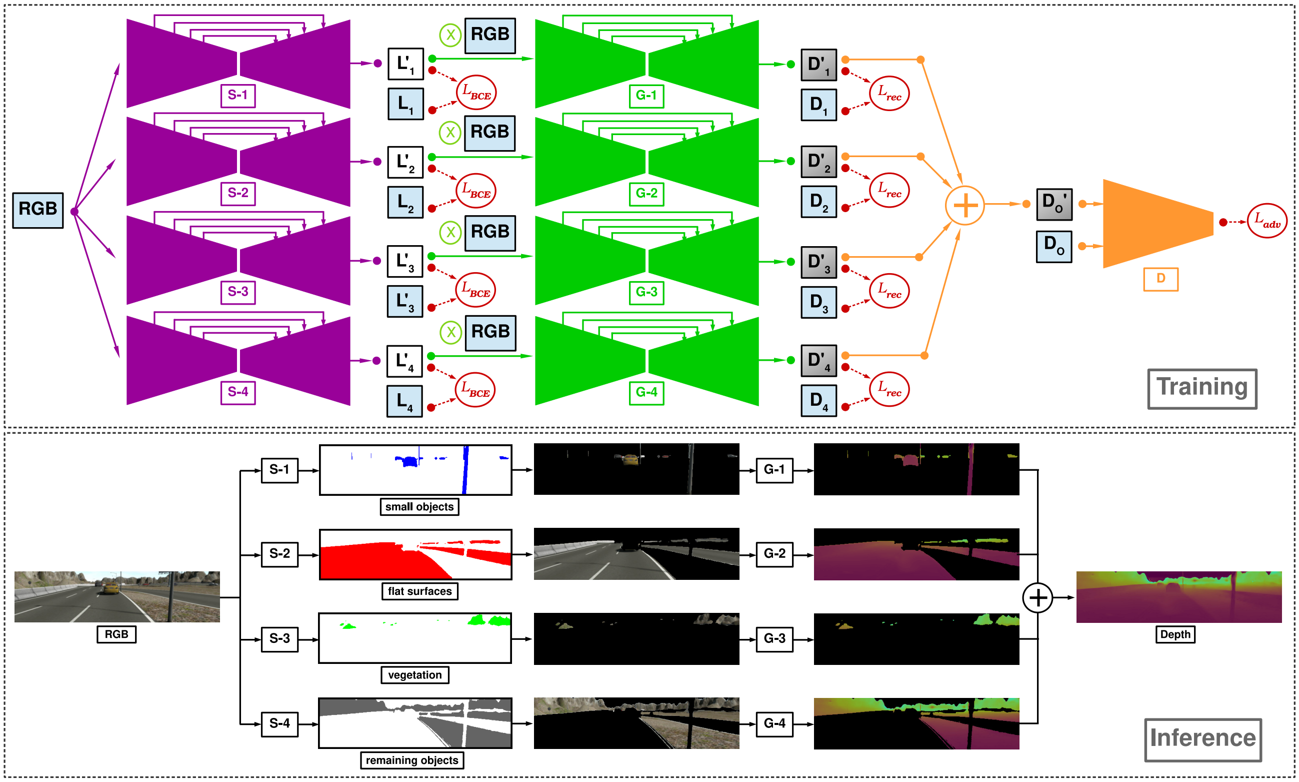

Fig. 2: Overall training (top) and inference (bottom) procedure of our model. The segmentation networks (S) are depicted in magenta, the depth generators (G) in green, and the discriminator (D) in orange. Loss functions are shown in red.

truth depth. The work in [16] presents a model that generates the right view from the left image used as the input while producing an intermediary disparity image. Similarly, [12] uses bilinear sampling [17] and a left/right consistency check incorporated into training for improved results.

In [11], depth and camera motion are predicted by training networks that estimate depth and pose, indirectly supervised via view synthesis. In [18], the model is trained using sparse ground truth depth and subsequently enforced within a stereo framework via an image alignment loss to output dense depth.

There are also supervised approaches [9, 19] that use synthetic images to produce depth outputs. Here, we also employ synthetic images [13] in a directly supervised training framework to perform the task of monocular depth estimation.

3. PROPOSED APPROACH

Our approach is designed to estimate depth for separate object groups that cover the entire scene when put together. Based on empirical analysis, we opt for decomposing any scene captured within an urban driving scenario into four object groups: (1) small and narrow foreground objects (e.g. pedestrians, road signs, cars) (2) flat surfaces (e.g. roads, buildings) (3) vegetation (e.g. trees, bushes) (4) background objects forming the remaining of the scene (other often unlabelled objects, e.g. a bench on the pavement).

For the sake of notation consistency, we will henceforth refer to the labels of this object groups as L_1, L_2, L_3 and L_4, respectively. Given an input image, each group is segmented using a separate segmentation network (S), the outputs of which are object groups that are subsequently employed to choose sections of the RGB image passed as inputs to depth generators (G), producing depth information for each object group (D_1, D_2, D_3, D_4). The entire model is trained as one entity, end to end (Figure 2). Such a training process is made possible using a synthetic dataset [13] in which both ground truth depth and pixel-wise segmentation labels are available for video sequences of urban driving scenarios.

3.1. Semantic Segmentation

For our segmentation networks (S), we opt for a simple and efficient fully-supervised training procedure. The RGB image is used as the input to all the networks and each network outputs class labels for its specific object group. A sigmoid function along with binary cross-entropy is used as the loss function for each network:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)), \quad (1)$$

where N denotes the number of pixels, y the ground truth label, and p is the predicted probability. This loss function is calculated for each of the four segmentation networks with the overall segmentation loss as follows:

Method	Error				Accuracy
	Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25^3$
Direct	0.861	1.894	7.012	0.488	0.683
Implicit	0.404	1.548	6.324	0.308	0.838
Explicit	**0.286**	**1.432**	**6.122**	**0.272**	**0.902**

Table 1: Comparing the performance of a single network estimating full depth (Direct), depth generators implicitly considering segments (Implicit) and the full approach (Explicit).

$$\mathcal{L}_{\text{seg}} = \sum_{n=1}^{4} \lambda_{\text{BCE}_n} \mathcal{L}_{\text{BCE}_n}, \tag{2}$$

where λ_{BCE} is the weighting coefficient empirically selected.

3.2. Monocular Depth Estimation

We consider monocular depth estimation as a supervised image-to-image mapping problem, wherein an input RGB image is translated into a depth image. More formally, a depth generator network (G) approximates a mapping function that takes as its input an RGB image x and outputs a depth image y, $G : x \to y$. As a result, the objective of the network should be to produce depth outputs that are as similar to the ground truth as possible. The most efficient and reliable solution is to minimize the Euclidean distance between the pixel values of the output, $G(x)$, and the ground truth depth, y. This simple reconstruction mechanism forces the model to generate images that are structurally and contextually close to the ground truth. In our approach, this reconstruction loss for the depth generator network G_1 (responsible for estimating the depth for small foreground object) is:

$$\mathcal{L}_{\text{rec}} = ||G_1(S_1(x) \times x) - (S_1(x) \times y)||_1, \tag{3}$$

where x is the RGB view of the entire scene, S_1 the segmentation network (Section 3.1), and y denotes the ground truth depth. This loss function is similarly calculated for each of the four segmentation networks, with the overall segmentation loss as follows:

$$\mathcal{L}_{\text{depth}} = \sum_{n=1}^{4} \lambda_{\text{rec}_n} \mathcal{L}_{\text{rec}_n}, \tag{4}$$

where λ_{rec} is the weighting coefficient empirically selected.

Once depth is estimated for each individual segment, the final depth output is obtained by simply summing the partial depth images generated for each segment. The overall depth is thus created as follows:

$$D_O = \sum_{n=1}^{4} D_n. \tag{5}$$

With a simple linear operation such as above, there is always the possibility of undesirable artefacts, such as stitching effects, over-saturation of depth values and depth inhomogeneity, being introduced in the results. To prevent such issues, we introduce an adversarial loss component that ensures the overall consistency of the final depth image. The result of the summation operation is used as the input to a discriminator, inspired by the work in [20], along with the ground depth overall depth. The gradients from this discriminator are used in the overall training of the entire model. Given the RGB input, x, our overall model (G_O) generates the entire scene depth output, $G_O(x) = \tilde{y}$ (result of Eqn. 5). Our discriminator (D) is adversarially trained to distinguish fake depth images produced by the model, $\tilde{y}$, from ground truth depth, y. The adversarial loss is thus as follows:

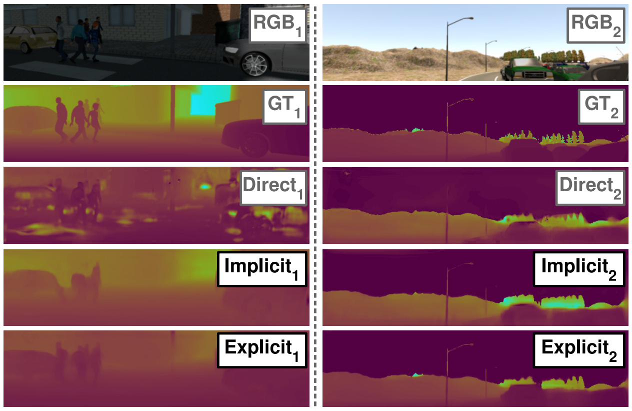

Fig. 3: Qualitative comparison of a single network directly estimating depth (Direct), depth generators implicitly considering segments (Implicit) and the full approach (Explicit).

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \min_{G_O} \max_{D} & \mathop{\mathbb{E}}_{x,y \sim \mathbb{P}_d(x,y)} [\log D(x,y)] + \\ & \mathop{\mathbb{E}}_{x \sim \mathbb{P}_d(x)} [\log(1 - D(x, G_O(x)))],\end{aligned} \tag{6}$$

where $\mathbb{P}_d$ is the data distribution defined by $\tilde{y} = G_O(x)$, x the input and y the ground truth. This loss ensures the fidelity of the overall depth output with no undesirable artefacts. Subsequently, the entirety of the model is trained end to end as one entity with the overall loss function as follows:

$$\mathcal{L} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{adv}}. \tag{7}$$

3.3. Implementation Details

Training data is composed of a large corpus of synthetic images [13] consisting of RGB, depth and pixel-wise class labels. For the sake of consistency, all the segmentation and depth generator networks follow a similar encoder-decoder architecture, containing skip connections [21] between every pair of corresponding layers in the encoder and the decoder. Our discriminator follows the architecture of [22] and, similar to our segmentation and depth generator networks, uses convolution-BatchNorm-leaky ReLU ($\textit{slope} = 0.2$) modules.

All implementation is done in *PyTorch* [23], with Adam [24] providing the best optimization ($\beta_1 = 0.5$, $\beta_2 = 0.999$, $\alpha = 0.0001$). The weighting coefficients in the loss function are empirically chosen to be $\lambda_{\text{BCE}_1} = 100$, $\lambda_{\text{BCE}_2} = 1$, $\lambda_{\text{BCE}_3} = 10$, $\lambda_{\text{BCE}_4} = 1$, $\lambda_{\text{rec}_1} = 100$, $\lambda_{\text{rec}_2} = \lambda_{\text{rec}_3} = \lambda_{\text{rec}_4} = 10$.

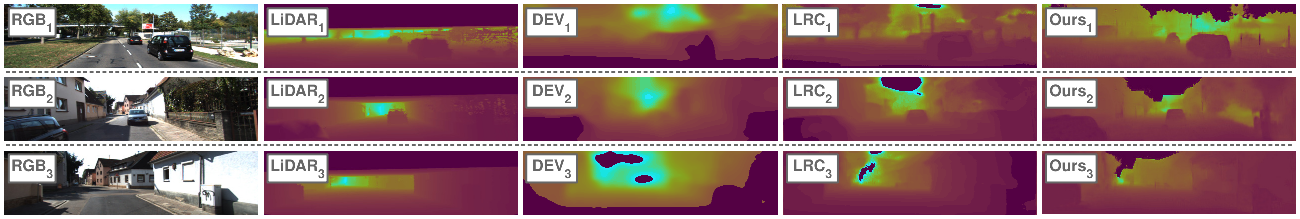

Fig. 4: Comparing the results of the approach against [11, 12]. Images have been adjusted for better visualization. **DEV:** Depth and Ego-motion from Video [11]; **LRC:** Left-Right Consistency [12].

4. EXPERIMENTAL RESULTS

We evaluate our approach using ablation studies and both qualitative and quantitative comparisons with contemporary methods applied to publicly available datasets [13, 25].

As a crucial part of our work, we attempt to demonstrate that the complexity introduced within the training and inference procedure is integral to the overall performance. Consequently, we simplify the model to measure the improvements made to the approach when an explicit segmentation step is performed before segment-wise depth estimation.

As a part of our ablation studies, we train a single network to carry out global scene depth estimation directly without the influence of any segment-wise segmentation. Secondly, we remove the explicit segmentation step and train the depth generator networks to implicitly perform segment-wise depth estimation by changing the ground truth they attempt to regress to. In essence, the ground truth class labels are used to guide each depth generator to learn to estimate depth for its specific segment only. Essentially, Eqn. 3 is changed as follows:

$$\mathcal{L}_{\text{rec}} = ||G_1(x) - (L_1 \times y)||_1, \qquad (8)$$

where x is the entire RGB image, L_1 the ground truth labels for small foreground objects and y denotes the ground truth scene depth. The same procedure is used for G_2, G_3 and G_4. The three resulting models (direct, implicit and explicit) are tested on randomly selected synthetic images [13].

As seen in Table 1, the implicit model produces promising results, offering potential future research directions. However, the explicit model outperforms the direct and implicit models, demonstrating the positive influence of complexity over its performance. Likewise, Figure 3 also illustrates the superiority of our full approach, specifically, when it comes to small and narrow objects in the scene.

Our approach is also evaluated against contemporary seminal depth estimation approaches [10, 11, 12, 26]. Following the conventions of the literature, we use the data split suggested in [10] for testing. In our assessments, the generated depth is corrected for the differences in focal length between the training [13] and testing data [25]. As seen in Table 2, our approach outperforms [10, 11, 26] across all metrics and stays very competitive with [12], with superior performance across most metrics. All of these comparators are trained and tested on the *same* dataset [25] while our approach is trained on [13] *without domain adaptation* and has not seen a single image from [25]. This is indeed very impressive and points to the generalization capabilities of our approach. We also experimented with re-training the comparators [11, 12] using the synthetic dataset, and as expected, our approach offers far superior performance. These results are not included in the interest of space (for brevity - RSME: 7.62 [11], 7.24 [12], 6.12 [ours]). As seen in Figure 4, the visual quality of the results of our approach exceeds that of the comparators. For better quality results, we kindly invite the reader to view the supplementary *video* material accompanying the paper.

Method		Error			Accuracy	
		Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25^3$
Train Set Mean	[25]	0.403	0.530	8.709	0.488	0.878
Eigen et al.	[10]	0.203	1.548	6.307	0.308	0.958
Liu et al.	[26]	0.202	1.614	6.523	0.308	0.965
Zhou et al.	[11]	0.208	1.768	6.856	0.283	0.957
Godard et al.	[12]	**0.148**	1.344	5.927	**0.247**	0.964
Our Approach		0.168	**1.338**	**5.702**	0.252	**0.968**

Table 2: Comparing our approach against [10, 26, 11, 12] using the split in [10]. Comparators are trained and tested on [25] while our approach is trained on [13] and tested on [25].

5. CONCLUSION AND FUTURE WORK

We propose a novel monocular depth estimation approach that utilizes a jointly-trained pixel-wise semantic understanding model to estimate depth for groups of objects. Estimating depth for scene segments leads to simpler learning objectives, which results in higher accuracy and better generalization capabilities. Extensive experimentation demonstrates the efficacy of the proposed approach compared to some of the best-performing techniques within the literature. Ablation studies also demonstrate that removing the segmentation step prior to depth prediction and having the depth generators implicitly perform segment-wise depth estimation can also produce promising results. This is important for future research since the complexity of the explicit model makes it intractable for real-time applications (112 ms per one forward pass), while the implicit model is much closer to the real time (61 ms). Furthermore, since the model is trained on synthetic data, the use of domain adaptation may also significantly improve the performance of the approach on real-world data.

For more information and results, we kindly invite the readers to refer to the ***video****: https://vimeo.com/336285373.*

6. REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Computer Vision*, vol. 47, pp. 7–42, 2002.

[2] L. Ding and G. Sharma, "Fusing structure from motion and Lidar for dense accurate depth map estimation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 1283–1287.

[3] M.W. Tao, P.P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1940–1948.

[4] R.J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 191139, 1980.

[5] A. Atapour-Abarghouei and T.P. Breckon, "A comparative review of plausible hole filling strategies in the context of scene depth image completion," *Computers and Graphics*, vol. 72, pp. 39–58, 2018.

[6] A. Atapour-Abarghouei and T.P. Breckon, "Extended patch prioritization for depth filling within constrained exemplar-based RGB-D image completion," in *Int. Conf. Image Analysis and Recognition*, 2018, pp. 306–314.

[7] A. Atapour-Abarghouei and T.P. Breckon, "DepthComp: Real-time depth image completion based on prior semantic scene segmentation," in *British Machine Vision Conference*, 2017, pp. 1–13.

[8] A. Atapour-Abarghouei, G. Payen de La Garanderie, and T.P. Breckon, "Back to Butterworth - a Fourier basis for 3D surface relief hole filling within RGB-D imagery," in *Int. Conf. Pattern Recognition*, 2016, pp. 2813–2818.

[9] A. Atapour-Abarghouei and T.P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 2800–2810.

[10] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, 2014, pp. 2366–2374.

[11] T. Zhou, M. Brown, N. Snavely, and D.G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 6612–6619.

[12] C. Godard, O. Mac Aodha, and G.J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 6602–6611.

[13] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Int. Conf. 3D Vision*, 2016, pp. 239–248.

[15] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1119–1127.

[16] J. Xie, R. Girshick, and A. Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Euro. Conf. Computer Vision*, 2016, pp. 842–857.

[17] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[18] Y. Kuznietsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 6647–6655.

[19] A. Atapour-Abarghouei and T.P. Breckon, "Veritatem dies aperit-temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, , A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Advances in Neural Information Processing Systems*, 2017, pp. 1–4.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Representations*, 2014, pp. 1–15.

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Robotics Research*, pp. 1231–1237, 2013.

[26] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, 2016.