Depth Estimation from Monocular Images Using Dilated Convolution and Uncertainty Learning

Haojie Ma[1,2], Yinzhang Ding[1,2], Lianghao Wang[1,2,3(✉)], Ming Zhang[1,2], and Dongxiao Li[1,2]

[1] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
wanglianghao@zju.edu.cn
[2] Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Hangzhou 310027, China
[3] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, People's Republic of China

Abstract. Depth cues are vital in many challenging computer vision tasks. In this paper, we address the problem of dense depth prediction from a single RGB image. Compared with stereo depth estimation, sensing the depth of a scene from monocular images is much more difficult and ambiguous because the epipolar geometry constraints cannot be exploited. In addition, the value of the scale is often unknown in monocular depth prediction. To facilitate an accurate single-view depth prediction, we introduce dilated convolution to capture multi-scale contextual information and then present a deep convolutional neural network. To improve the robustness of the system, we estimate the uncertainty of noisy data by modelling such uncertainty in a new loss function. The experiment results show that the proposed approach outperforms the previous state-of-the-art methods in depth estimation tasks.

Keywords: Depth estimation · Dilated convolution
Convolutional neural network · Uncertainty

1 Introduction

Depth estimation has been investigated for a long time because of its vital role in computer vision. Some studies have proven that accurate depth information is useful for various existing challenging tasks, such as image segmentation [1], 3D reconstruction [2], human pose estimation [3], and counter detection [5]. Humans can effectively predict monocular depth by using their past visual experiences to structurally understand their world and may even utilize such knowledge in unfamiliar environments. However, monocular depth prediction remains a difficult problem for computer vision systems due to the lack of reliable depth cues.

R. Hong et al. (Eds.): PCM 2018, LNCS 11165, pp. 13–23, 2018.
https://doi.org/10.1007/978-3-030-00767-6_2

Many studies have investigated the use of image correspondences that are included in stereo image pairs [6]. In the case of stereo images, depth estimation can be addressed when the correspondence between the points in the left and right parts of images is established. Many studies have also explored the method of motion [7], which initially estimates the camera pose based on the change in motion in video sequences and then recovers the depth via triangulation. Obtaining a sufficient number of point correspondences plays a key role in the aforementioned methods. These correspondences are often found by using the local feature selection and matching techniques. However, the feature-based method usually fails in the absence of texture and the occurrence of occlusion. Owing to the recent advancements in depth sensors, directly measuring depth has recently become affordable and achievable, but these sensors have their own limitations in practice. For instance, Microsoft Kinect is widely used indoors for acquiring RGB-D images but is limited by short measurement distance and large power consumption. When working outdoors, LiDAR and relatively cheaper millimeter wave radars are mainly used to capture depth data. However, these collected data are always sparse and noisy. Accordingly, there has always been a strong interest in accurate depth estimation from a single image. Recently, CNNs [8] with powerful feature representation capabilities have been widely used for single-view depth estimation by learning the implicit relationship between an RGB image and the depth map. Despite increasing the complexity of the task, the outputs of deep-learning-based approaches [13–17] showed significant improvements over those of traditional techniques [10–12] on public datasets.

In this work, we exploit CNNs to learn strong features for recovering the depth map from a single still image. Given that applying downsampling, upsampling, or deconvolution in a fully convolution network may result in the loss of many cues in the image boundary for pixel-level regression tasks, we introduce the dilated convolution [9] to learn multi-scale information from a single scale image input. Long skip connections are also applied to combine the abstract features with image features. To achieve more robust prediction, we further model the uncertainty in computer vision and proposed a novel loss function to measure such uncertainty during training without labels. The experimental results demonstrate that our proposed method outperform state-of-the-art approaches on standard benchmark datasets [2,21].

2 Related Work

Previous studies have often used probabilistic graphic model and have usually relied on hand-crafted features. Saxena et al. [10] proposed a superpixel-based method for inferring depth from a single still image and applied a multi-scale MRF to incorporate local and global features. Ladicky et al. [11] introduced semantic labels on the depth to learn a highly discriminative classifier. Karsch et al. [12] proposed a non-parametric approach for automatically generating depth. In this approach, the GIST features were initially extracted for the input image and other images in the database, then several candidate depths that

correspond to the candidate images were selected before conducting warping and optimization procedures.

Recent studies have employed CNNs to solve the depth prediction problem. Eigen et al. [13] utilized two network stacks to regress depth. The first local network makes a coarse depth prediction for the global contents, while the other network refines the prediction locally. Liu et al. [15] combined CNN with continuous CRF in a unified framework. Wang et al. [16] jointly addressed depth prediction and semantic segmentation by using a common CNN. They proposed a two-layer hierarchical CRF model to refine the coarse network output. Laina et al. [17] proposed a fully convolutional network and introduced a robust loss function called berHu loss. Some unsupervised approaches have also been introduced recently to address the challenges in obtaining a large number of dense and reliable depth labels, especially in outdoor scenes. Garg et al. [18] treated depth estimation as an image reconstruction problem and proposed the use of photometric loss. Given that the loss is not completely differentiable, they performed a first-order Taylor expansion to linearize the results in warp images. Based on [18], Godard et al. [19] considered the left-right disparity consistency constraint and dealt with the warp image by bilinear interpolation.

In this paper, we construct a fully convolutional network for monocular depth prediction. To maintain additional feature information, we apply dilated convolution to enlarge the receptive field without reducing the resolution of the feature maps. Afterward, we implement low-level and high-level information fusion by using long skip connections. Unlike the previous CNNs that are unable to represent, or model uncertainty as probability distributions by using CRF, our network can accurately estimate uncertainty as the model attenuation.

3 Approach

3.1 Network Architecture

We adopt a deep fully convolutional network with an encoder-decoder architecture for single-view depth estimation (see Fig. 1). This network is constructed based on ResNet [4], which performs well in image classification. We remove the last average pooling layer and fully connected layer of the original version. In this way, we discard most of the network parameters and successfully train our model on the current hardware. As low-resolution feature maps contain less boundary information, we employ dilated convolution to expand the receptive field while maintaining the features within an appropriate size. The key components of the decoder part are the two up-sampling layers that are used to recover image resolution. To achieve a higher depth accuracy, we choose the up-projection module proposed in [17] as our up-sampling layer. This module comprises an unpooling layer (which increases the spatial resolution of the feature map) and two convolution layers with residual learning. We concatenate the corresponding feature maps from the encoder and decoder parts by skip connections. Eventually, a convolution is applied to generate the depth prediction.

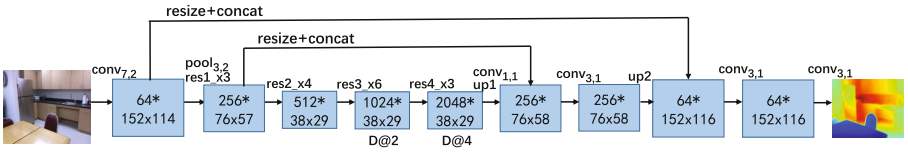

Fig. 1. Model architecture. We use $\text{conv}_{\text{n,s}}$ to denote a n × n convolution with stride s, and same notation is employed to max pooling $\text{pool}_{\text{n,s}}$. Let k* be feature maps and @r be the dilation rate. Residual blocks (res1, res2, res3, res4) consist of three convolutions with kernel size 1 × 1, 3 × 3, 1 × 1 and xm is the number of blocks. We replace the convolution by dilated convolution in res3_x6 and res4_x3.

Dilated Convolution. Dilated convolution has been recently proposed to overcome the reduced feature resolution problem caused by the successive pooling and down-sampling layers. Dilated convolution is a regular convolution with a kernel that is dilated by inserting zeros between non-zero values. Compared with standard convolution, dilated convolution can effectively increase the receptive field without increasing the number of parameters. Multi-scale contextual information is also extracted from the original resolution. A dilated convolution is defined as

$$(F *_n k)(\mathbf{p}) = \sum_{\mathbf{s}+n\mathbf{t}=\mathbf{p}} F(\mathbf{s})k(\mathbf{t}), \tag{1}$$

where $F : \mathbb{Z}^2 \rightarrow \mathbb{R}$ is a discrete function, n is the dilation rate, $*_n$ is an n-dilated convolution, and $\Omega_{\text{r}} = [-r, r]^2 \cap \mathbb{Z}^2$. Let $\text{k} : \Omega_{\text{r}} \rightarrow \mathbb{R}$ be a discrete filter of size $(2\text{r} + 1)^2$.

3.2 Loss Function with Uncertainty Learning

Two major types of uncertainty can be modeled in deep learning. First, epistemic uncertainty, also known as systematic uncertainty, describes the uncertainty over the model parameters. Second, aleatoric uncertainty, also called statistical uncertainty, represents the inherent noise in the inputs and cannot be decreased no matter how much data are provided. We specifically focus on modelling aleatoric uncertainty, since epistemic uncertainty can be mostly eliminated by using large amounts of data.

To learn aleatoric uncertainty, we measure the variance of noise from the input RGB images. Compared with previous CNNs for depth prediction where the noise parameter σ is replaced by a fixed weight decay, our scheme assumes that the noise is variable for different inputs, since the depth for textureless regions is highly ambiguous. For a predicted depth map $\widetilde{y}$ and the corresponding ground truth y, the variance is learned as loss attenuation and we define the new loss as

$$\mathcal{L} = \frac{1}{n}\sum_{i}^{n} \widetilde{\sigma}_i^{-2} \|y_i - \widetilde{y}_i\|_2^2 + \log \widetilde{\sigma}_i^2, \tag{2}$$

where i indexes the n pixels over the image, and $\widetilde{\sigma}_i^2$ denotes the variance for pixel i. This loss consists of two components: a residual regression term and an uncertainty regularization term.

In practice, we predict $\widetilde{s}_i := \log \widetilde{\sigma}_i^2$ and

$$\mathcal{L} = \frac{1}{n} \sum_{i}^{n} \exp\left(-\widetilde{s}_i\right) \|y_i - \widetilde{y}_i\|_2^2 + \widetilde{s}_i. \tag{3}$$

The loss in Eq. 3 has a better numerical stability than that in Eq. 2 by avoiding division by zero.

The $\mathcal{L}_1 = \|y - \widetilde{y}\|_1$, $\mathcal{L}_2 = \|y - \widetilde{y}\|_2^2$, and berHu loss [17] were separately tested in the experiment, and the results revealed that $\mathcal{L}_1$ outperformed the others in estimating monocular depth. An explicit quantitative analysis is shown in Sect. 4. Therefore, we adopt the $\mathcal{L}_1$ norm instead of $\mathcal{L}_2$ norm as the residual term described in Eq. 3 during training.

4 Experiments

In this section, we perform with a quantitative analysis to test our proposed method. and then compare the performance of this method with other start-of-the-art models on two popular datasets, namely, NYU Depth v2 [21] and Make3D [2].

4.1 Experimental Setup

For the following analysis and evaluation, we implement our architecture by using Tensorflow, and train on a single NVIDIA GeForce GTX 1080Ti with 11GB memory. The weights of the network are initialed by the ResNet-50 model that is pre-trained on ImageNet data [23]. In all experiments, batch size and weight decay are set to 16 and 10^{-4}, respectively. We train the network for approximately 15 epochs on NYU Depth v2 and 20 epochs on Make3D. The starting learning rate is 0.001 and halved every 5 epochs. As for the initial values of the log variances, we set $s = 0.0$.

4.2 Datasets

The NYU Depth v2 dataset [21] contains 120 K unique RGB-D images taken from 464 different indoor scenes with a Kinect camera. We use 249 scenes for training and the other 215 scenes for testing according to the official train/test split. We sample equally spaced frames from each raw training sequence and obtain approximately 12 K RGB-D pairs. The missing depth values are filled in by using the toolbox provided by Silberman et al. [21]. To increase the size and variability of the training set, we perform a data augmentation similar to that in [13], and get roughly 96 K pairs. Following [17], the original frames are downsampled by half and then center-cropped to 304×228. For testing, we use

Table 1. Quantitative analysis of proposed architectures on the official test set of NYU Depth v2. For rel, rms, and $\log_{10}$, a lower is better, for δ_1, δ_2 and δ_3, a higher is better. Results in bold are best.

Architecture	Loss	rel	rms	$\log_{10}$	δ_1	δ_2	δ_3
Baseline	berHu	0.128	0.573	0.055	0.801	0.950	0.985
Ours (dilated convolution)	berHu	0.122	0.565	0.052	0.805	0.953	0.986
Ours (long skip connections)	berHu	0.118	0.560	0.050	0.814	0.955	0.988
Ours (full)	berHu	0.115	0.556	0.049	0.816	0.956	0.988
Ours (full)	$\mathcal{L}_2$	0.130	0.572	0.054	0.799	0.950	0.985
Ours (full)	$\mathcal{L}_1$	0.113	0.553	0.049	0.817	0.956	0.988
Ours (full)	$\mathcal{L}_1$+uncertainty	**0.110**	**0.552**	**0.048**	**0.820**	**0.958**	**0.989**

Table 2. Performance comparison with state-of-the-art methods on the NYU Depth v2 dataset. The values are originally reported by the authors in their respective papers

Method	rel	rms	$\log_{10}$	rms(log)	δ_1	δ_2	δ_3
Li et al. [26]	0.232	0.821	0.094	-	0.621	0.886	0.968
Liu et al. [15]	0.230	0.824	0.095	-	0.614	0.883	0.971
Wang et al. [16]	0.220	0.745	0.094	0.262	0.605	0.890	0.970
Eigen et al. [13]	0.215	0.907	-	0.285	0.611	0.887	0.971
Roy et al. [24]	0.187	0.744	0.078	-	-	-	-
Eigen and Fergus [14]	0.158	0.641	-	0.214	0.769	0.950	0.988
Laina et al. [17]	0.127	0.573	0.055	0.195	0.811	0.953	0.988
Xu et al. [25]	0.121	0.586	0.052	-	0.811	0.954	0.987
Ours	**0.110**	**0.550**	**0.048**	**0.173**	**0.820**	**0.958**	**0.989**

Table 3. Performance comparison with state-of-the-art methods on the Make3D dataset. The values are originally reported by the authors in their respective papers

Method	rel	rms	$\log_{10}$
Liu et al. [20]	0.335	9.49	0.137
Liu et al. [15]	0.314	8.60	0.119
Li et al. [26]	0.278	7.19	0.092
Laina et al. [17]	0.176	4.46	0.072
Xu et al. [25]	0.184	4.38	0.065
Ours	**0.165**	**4.35**	**0.063**

654 images from the labeled part of the dataset. The predictions are resized to 640 × 480 via bilinear interpolation to evaluate the performance of the model.

Make3D [2] is an outdoor scene dataset that consists of 534 RGB-D pairs, which are separated into 400 training images and 134 test images. Due to the limitations of the hardware, we resize the original images from 1704 × 2272 to 345 × 460. During training, RGB images are halved again. We also augment the training data with offline transformations and obtain about 15k samples. Given that the laser scanner has a maximum range of 81 m, we only compute the error for those pixels with a ground-truth depth less than 70 m.

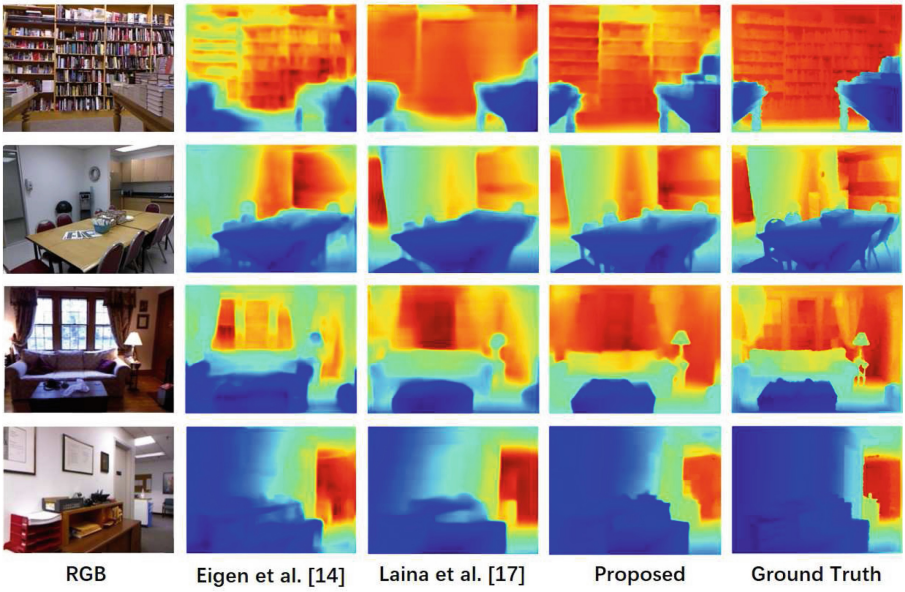

Fig. 2. Qualitative results on the NYU Depth v2 dataset. For fair comparison, all depth predictions shown in color are scaled equally (blue is close and red is far). (Color figure online)

4.3 Evaluation Metrics

To objectively evaluate the performance of our depth estimation model, we use the following metrics:

- mean absolute relative error (rel): $\frac{1}{N}\sum_i \frac{|y_i - \widetilde{y}_i|}{y_i}$;
- root mean square error (rms): $\sqrt{\frac{1}{N}\sum_i (y_i - \widetilde{y}_i)^2}$;
- root mean square log error (rms(log)): $\sqrt{\frac{1}{N}\sum_i (\log y_i - \log \widetilde{y}_i)^2}$;

– mean $\log_{10}$ error ($\log_{10}$): $\frac{1}{N}\sum_i |\log y_i - \log \widetilde{y}_i|$;
– δ_j: percentage of $\widetilde{y}_i$ s.t. $\max(\frac{\widetilde{y}_i}{y_i}, \frac{y_i}{\widetilde{y}_i}) < 1.25^j$.

Where y_i and $\widetilde{y}_i$ are the ground-truth depth and predicted depth at pixel indexed by i, and N is the number of pixels.

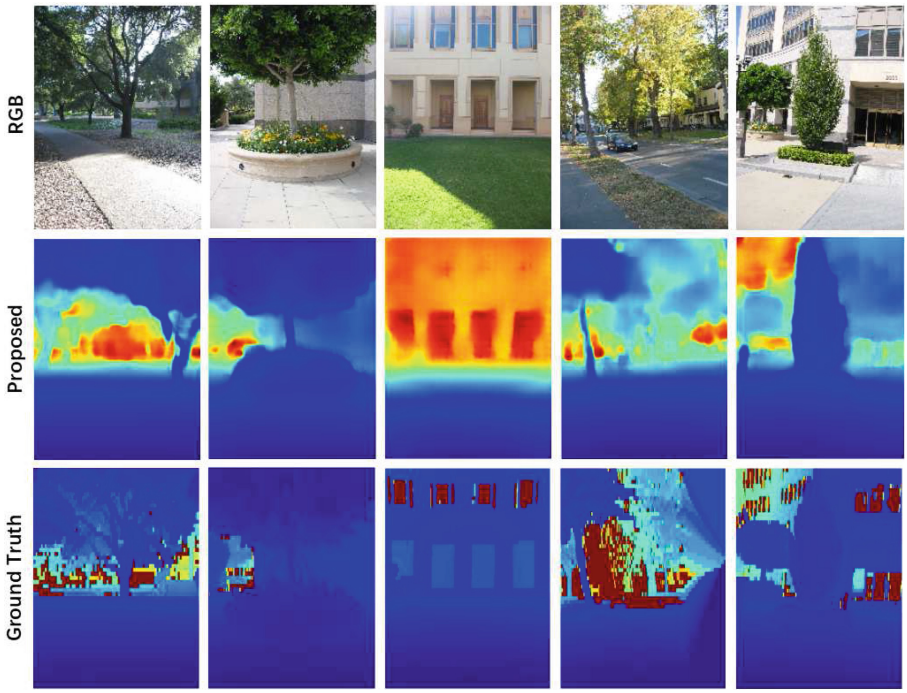

Fig. 3. Qualitative results on the Make3D dataset. We estimate depth for all pixels in the color maps.

4.4 Results

Architecture Evaluation. In this section, we analyze the effects of different architectures and loss functions on depth estimation performance. The quantitative results are shown in Table 1. For an ablation study, we train a baseline network composed of ResNet and up-sampling layers on the NYU Depth v2 dataset (row 1 in Table 1). To demonstrate the effectiveness of dilated convolution, we replace the last two down-sampling regular convolutions with 3×3 dilated convolutions and obtain better results (row 2 in Table 1). Table 1 also shows that long skip connections added to the baseline network can significantly improve the depth estimation performance (row 3, 4). Obviously, multi-scale contextual information fusion is beneficial to depth regression. Moreover, we compare the

$\mathcal{L}_1$, $\mathcal{L}_2$, and berHu loss functions with our proposed architecture. Rows 4 to 6 in Table 1 show that $\mathcal{L}_1$ is greater than both $\mathcal{L}_2$ and berHu. We further combine the $\mathcal{L}_1$ loss with uncertainty learning and achieve the best result as shown in row 7.

Comparison with the State-of-the-Art. Table 2 compares the performance of our method and that of several state-of-the-art approaches on the NYU Depth v2 dataset. Due to the use of dilated convolution, long skip connections, and heteroscedastic uncertainty, our method outperforms other approaches on all metrics. The quantitative results in Fig. 2 illustrate that the proposed model accurately estimates the depth in textureless regions (e.g., walls) and image edges. To demonstrate the generalization ability of the proposed model, we also compare its performance with that of previous related works on the Make3d dataset. Table 3 shows that our model outperforms the other state-of-the-art methods. Additional quantitative examples are provided in Fig. 3.

5 Conclusion

In this paper, we propose a novel approach for solving the monocular depth estimation problem. We introduce a deep residual network with dilated convolution and long skip connections that can aggregate multi-scale contextual information and generate a detailed depth map. By modelling the input-dependent aleatoric uncertainty as learned attenuation, we reduce the effect of noisy data and improve the accuracy of the depth estimation. The experimental results on two benchmark datasets demonstrate that our proposed method outperforms the other state-of-the-art approaches.

Depth information is beneficial for addressing various computer vision problems. In our future work, we plan to examine the application of our depth model to other tasks, such as object detection, semantic segmentation, and simultaneous localization and mapping.

Acknowledgments. This work was supported in part by Zhejiang Provincial Natural Science Foundation of China (Grant No. LY18F010004).

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. IEEE, Boston (2015)
2. Saxena, A., Sun, M., Ng, A.Y.: Make3D: learning 3D scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 824–840 (2009)
3. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The Vitruvian manifold: inferring dense correspondences for one-shot human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 103–110. IEEE, Providence (2012)

4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Las Vegas (2016)
5. Bertasius, G., Shi, J., Torresani, L.: DeepEdge: a multi-scale bifurcated deep network for top-down contour detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4380–4389. IEEE, Boston (2015)
6. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. **47**(1–3), 7–42 (2002)
7. Szeliski, R.: Structure from motion. Computer Vision. Texts in Computer Science, pp. 303–334. Springer, London (2011). https://doi.org/10.1007/978-1-84882-935-7
8. LeCun, Y., et al.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
9. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations, Caribe Hilton, Puerto Rico (2016)
10. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: International Conference on Neural Information Processing Systems, pp. 1161–1168. MIT Press, Vancouver (2005)
11. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 89–96. IEEE, Columbus (2014)
12. Karsch, K., Liu, C., Kang, S.B.: Depth extraction from video using non-parametric sampling. In: Fitzgibbon, A., et al. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 775–788. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_56
13. Eigen, D., Puhrsch, C., Fergus, R.: Prediction from a single image using a multi-scale deep network. In: International Conference on Neural Information Processing Systems, pp. 2366–2374. MIT Press, Montreal (2014)
14. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2650–2658. IEEE, Santiago (2015)
15. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5162–5170. IEEE, Boston (2015)
16. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2800–2809. IEEE, Boston (2015)
17. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248. IEEE, Stanford (2016)
18. Garg, R., Vijay Kumar, B.G., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
19. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6602–6611. IEEE, Honolulu (2017)
20. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723. IEEE, Columbus (2014)

21. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
22. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: a deep neural network architecture for real-time semantic segmentation. CoRR abs/1606.02147 (2016). http://arxiv.org/abs/1606.02147
23. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
24. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5506–5514. IEEE, Las Vegas (2016)
25. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 161–169. IEEE, Honolulu (2017)
26. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1119–1127. IEEE, Boston (2015)