17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

HMA-Depth: A New Monocular Depth Estimation Model Using Hierarchical Multi-Scale Attention

Zhaofeng Niu, Yuichiro Fujimoto, Masayuki Kanbara, Hirokazu Kato
Nara Institute of Science and Technology
{niu.zhaofeng.mv7, yfujimoto, kanbara, kato}@is.naist.jp

Abstract

Monocular depth estimation is an essential technique for tasks like 3D reconstruction. Although many works have emerged in recent years, they can be improved by better utilizing the multi-scale information of the input images, which is proved to be one of the keys in generating high-quality depth estimations. In this paper, we propose a new monocular depth estimation method named HMA-Depth, in which we follow the encoder-decoder scheme and combine several techniques such as skip connections and the atrous spatial pyramid pooling. To obtain more precise local information from the image while keeping a good understanding of the global context, a hierarchical multi-scale attention module is adopted and its outputs are combined to generate the final output that is with both good details and good overall accuracy. Experimental results on two commonly-used datasets prove that HMA-Depth can outperform the existing approaches. Code is available[1]*.*

1 Introduction

Depth sensing is an important technique for various applications [1, 2, 3], such as 3D reconstruction, autonomous driving, augmented reality, etc. Although there have existed various types of depth sensors like the structured-light 3D scanner and the time-of-flight camera, they have the following drawbacks [4]. Firstly, the resolution and sensing range of the existing 3D sensors are very limited. Secondly, 3D sensors usually cost significantly more than 2D cameras. Thirdly, 3D sensors also cause higher power consumption, which is a big concern for mobile devices. Therefore, to overcome these limitations, monocular depth estimation has drawn a lot of attention.

Monocular depth estimation is a process that obtaining the depth map from a single 2D image. monocular depth estimation is a very challenging task [5, 6], since one 2D image could be matched with infinite 3D scenes. However, with the rapid development of deep learning theories and convolutional neural networks (CNNs) in recent years, many encouraging works [7, 8, 9, 10] have emerged, showing greatly-improved results on mainstream datasets (*e.g.*, KITTI [11] and NYU V2 [12]).

[1] https://github.com/saranew/HMADepth

In the encoder-decoder-based computer vision tasks, there usually is a trade-off between preserving the fine details and achieving a good understanding of the global context [13, 14]. Due to the model structure and the mechanisms of convolutions, CNNs are good at keeping local information while are relatively weak at extracting global knowledge. Therefore, when we need a model that can well analyze the relationships among all the objects in the image, which is necessary for depth estimation, we have to down-scale the input image to let the model better learn the overall information. However, at the same time, prediction with the down-scaled image will also lose some details that are too small to analyze. On the contrary, when fine details are required, we prefer the large-scaled image, which, however, often leads to poor overall accuracy. A common solution is to use the images with multiple scales and combine their predictions together [13, 15]. However, most of the existing methods simply use some operations like averaging or max pooling, which are actually combining good predictions with poorer ones and, therefore, are not theoretically optimal.

To address the aforementioned problems, we propose a new monocular depth estimation model called hierarchical multi-scale attention-based depth estimation network (HMA-Depth) in the paper. We follow the encoder-decoder scheme, which is commonly used in computer vision tasks. In addition, an atrous spatial pyramid pooling (ASPP) module [16], which uses convolutional kernels with different dilation rates, is adopted to improve the feature quality. To enable the multi-scale depth estimation, we upsample the initial feature to larger scales, for some of which we attach a pair of depth estimation head and attention generation head to the corresponding features. The depth head gives the depth map and the attention head is for choosing (using a weight map A in which every weight $A_i \in [0, 1]$) the preferred regions in the generated depth map. Inspired by some semantic segmentation methods [17, 14], we adopt a hierarchical design for the attention heads, in which $\sum_{A \in \mathcal{A}} A_i = 1$, where i is an arbitrary point on the attention map A and $\mathcal{A}$ is the whole attention map set. The final result of depth estimation is a weighted sum of all depth maps generated at different scales.

As mentioned above, depth estimation at different scales has different advantages and disadvantages. We notice that the attention maps can accurately pick up

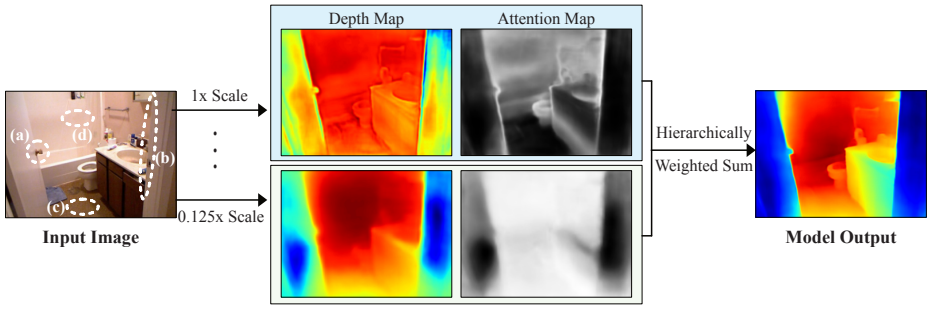

Figure 1. Depth estimation with hierarchical multi-scale attention. (a) and (b) are two local details that need prediction at large scale (1x), while (c) and (d) need a good overall understanding about the relationships among objects, where prediction at small scale (0.125×) is preferred.

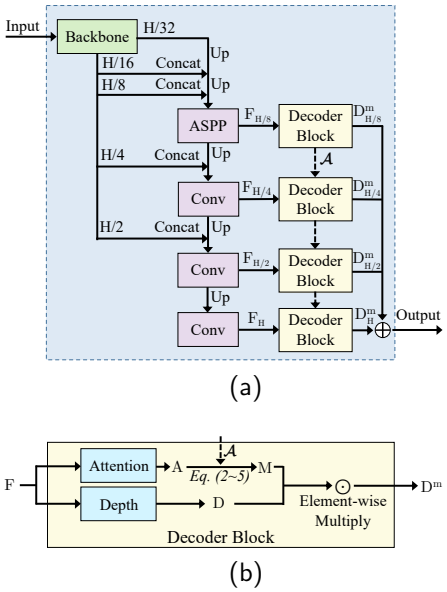

Figure 2. Network architecture. (a) The HMA-Depth model. (b) Details of the decoder block.

the advantages of the prediction at each scale. As shown in Fig. 1, the prediction at a larger scale (1x) is good at details, *e.g.*, (a) the handle and (b) the edge, while the prediction at the small scale (0.125×) is good at global understanding, *e.g.*, (c) the ground near the camera and (d) the wall far away. After the weighted sum, we can get a depth estimation with both details and overall accuracy.

In sum, our contributions are three-fold:

- We design a network that can generate features at different scales, each of which provides different information about the input image.
- A hierarchical multi-scale attention (HMA) module is adopted to generate depth estimations with both good local details and overall accuracy.
- An ablation study is conducted to find the optimal parameters for the HMA module.

2 Method

2.1 Network Architecture

As shown in Fig. 2 (a), the proposed HMA-Depth model follows the encoder-decoder scheme, in which the backbone module is the encoder part and the remaining modules are the decoder part. The input of the network is a single RGB image with original resolution $R = H \times W$. As the encoder part, we use a CNN model as the backbone to obtain the feature maps at different scales (the features generated by the last layer of the backbone as well as the intermediate features), of which the heights and widths are equally down-sampled and the resolutions are $H/32 \times W/32$, $H/16 \times W/16$, $H/8 \times W/8$, $H/4 \times W/4$, and $H/2 \times W/2$ (we will only use H/s to represent the scales for short and $s \in S = \{1, 2, 4, 8, 16, 32\}$), respectively. The direct output from the backbone will be up-sampled to larger scales and be concatenated with the skip connection from the intermediate features of the backbone. We use the bilinear interpolation and a 3×3 convolutional layer for the up-sampling process. Besides, an ASPP module is utilized for contextual information extraction. Similar to [15], we set the dilation rates of ASPP module as $r \in \{3, 6, 12, 18, 24\}$.

The output feature from ASPP will be further up-sampled several times. After each upsampling process, there is a convolutional module to process the features, which is a 3×3 convolutional layer for scale $H/4$ and $H/2$ (the first two *Conv*s in Fig. 2(a)) and a 1×1 convolutional layer for resolution H (the last *Conv*). For the feature H/s with $s \in S' = \{1, 2, 4, 8\}$, we attach the decoder block to analyze the scaled features, which can output the weighted depth map for each scale. We will explain the decoder block in detail in the next subsection. Finally, all four weighted depth maps are added together to generate the final output.

As for the loss function, we adopt the scale-invariant error proposed by Eigen et al [13], which calculates the error between a predicted depth map y and ground

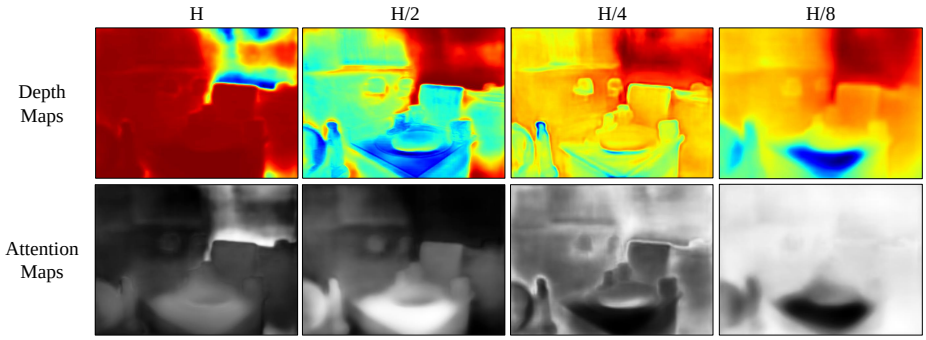

Figure 3. Depth and attention maps generated at different scales.

truth y^* as follows:

$$Loss = \frac{1}{n}\sum_i g_i^2 - \frac{\lambda}{n^2}(\sum_i g_i)^2 \quad (1)$$

where $g_i = \log y_i - \log y_i^*$ and $\lambda \in [0, 1]$; n indicates the number of pixels that have valid depth values. Similar to [15], we set $\lambda = 0.85$ to minimize the variance of the error.

2.2 Hierarchical Multi-Scale Attention

As shown in Fig. 2 (b), the decoder block can generate the depth map D and the attention map A, respectively with the depth head and attention head. The depth map is the depth estimation using the features at the corresponding scales, while the attention map can extract the preferred regions for each depth map, according to the image contents and the characteristics of the predictions at the corresponding scale. We use $D_{H/8}$, $D_{H/4}$, $D_{H/2}$ and D_H to represent the scaled depth maps, and $A_{H/8}$, $A_{H/4}$, $A_{H/2}$ and A_H to indicate the attention maps for the corresponding prediction. In our implementation, each depth head has two 3×3 and one 1×1 convolutional layers; each attention head has one 3×3 and one 1×1 convolutional layers.

A problem is how to combine the predictions at different scales. In our method, we adopt a hierarchical design to generate the weight masks, as shown below.

$$M_{H/8} = A_{H/8} \quad (2)$$
$$M_{H/4} = A_{H/4}(1 - A_{H/8}) \quad (3)$$
$$M_{H/2} = A_{H/2}(1 - A_{H/8})(1 - A_{H/4}) \quad (4)$$
$$M_H = (1 - A_{H/8})(1 - A_{H/4})(1 - A_{H/2}) \quad (5)$$

It can be seen that the prediction at each scale needs to pay different attention to the regions of the input image. Specifically, the sum of the masks is $\mathbf{1}$, which is a matrix with all elements equal to 1 (as shown in Fig. 3, where the white regions, *i.e.*, the areas with attention, are complementary among mask images and the sum of masks would be a whole white image, which means all areas in the image can be covered by amplifying the benefits of each scales).

Then the scaled depth maps and masks are element-wise multiplied into the weighted depth map D^m and the final depth map D_{final} is obtained by summing up the weighted depths of all predictions, which can be represented as follows:

$$D_{\text{final}} = \sum_{s\in S'} M_s \cdot D_s^m \quad (6)$$

In addition, we provide the visualization of the depth maps and attention maps for each scale in Fig. 3. We can see that the attention module can reasonably choose the preferred regions for each scale. A trend is that the model pays more attention to the depth values in small-scaled predictions, while relies on the large-scaled predictions for fine details, such as the edge and local information, which conforms to the intention of the network design.

3 Experiment

To have a complete evaluation of the HMA-Depth model, we conduct several different experiments on two commonly-used datasets, *i.e.*, KITTI dataset [11] and NYU V2 dataset [12], and the results are compared with the state-of-the-art approaches.

3.1 Implementation

PyTorch [18] is adopted to implement our network. The number of the epoch is set as 50 and the batch size is 16. We use a server with four NVIDIA V100 32G GPUs for all the experiments.

The backbone network is used to extract the dense feature. To prove the effectiveness of our network, we use multiple networks as the backbone network, including ResNet 50 [19], ResNeXt 50 [20], DenseNet 121 [21], and DenseNet 161 [21]. To avoid over-fitting, we adopt data augmentation techniques including random

Table 1. Quantitative results on KITTI dataset

Methods	Higher is better			Lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	RMSElog
Make3D.[22]	0.601	0.820	0.926	0.280	8.734	0.361
Eigen et al.[13]	0.702	0.898	0.967	0.203	6.307	0.282
Liu et al.[23]	0.680	0.898	0.967	0.201	6.471	0.273
Kuznietso et al [24]	0.862	0.960	0.986	0.113	4.621	0.189
Yin et al. [25]	0.938	0.990	**0.998**	0.072	3.258	0.117
DORN [5]	0.932	0.984	0.994	0.072	**2.727**	0.120
BTS-ResNet 50 [15]	0.950	0.991	**0.998**	0.062	2.878	0.101
BTS-DenseNet 161 [15]	0.952	0.992	**0.998**	0.062	2.871	0.094
Ours-ResNet 50	0.953	0.992	**0.998**	0.062	2.870	0.096
Ours-ResNeXt 50	0.951	0.992	**0.998**	0.062	2.867	0.094
Ours-DenseNet 121	0.952	0.991	**0.998**	0.063	2.874	0.096
Ours-DenseNet 161	**0.955**	**0.993**	**0.998**	**0.060**	2.850	**0.092**

Table 2. Quantitative results on NYU V2 dataset

Methods	Higher is better			Lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	log10
Make3D [22]	0.447	0.745	0.897	0.349	1.214	-
Wang et al. [26]	0.605	0.890	0.970	0.220	0.824	-
Liu et al. [23]	0.650	0.906	0.976	0.213	0.759	0.087
Eigen et al. [13]	0.769	0.950	0.988	0.158	0.641	-
Li et al. [27]	0.621	0.886	0.968	0.232	0.821	0.094
Laina et al. [28]	0.811	0.953	0.988	0.127	0.573	0.055
DORN [5]	0.828	0.965	0.992	0.115	0.509	0.051
Yin et al. [25]	0.875	0.976	0.994	**0.108**	0.416	0.048
BTS-ResNet 50 [15]	0.862	0.975	0.994	0.120	0.421	0.051
BTS-DenseNet 161 [15]	0.879	**0.980**	0.995	0.112	0.399	0.048
Ours-ResNet 50	0.866	0.977	0.994	0.118	0.417	0.050
Ours-ResNeXt 50	0.862	0.976	0.994	0.121	0.419	0.051
Ours-DenseNet 121	0.865	0.974	0.993	0.121	0.421	0.051
Ours-DenseNet 161	**0.882**	**0.980**	**0.996**	0.110	**0.394**	**0.047**

Table 3. Ablation results

Methods	Higher is better			Lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	log10
Base	**0.866**	**0.977**	**0.994**	**0.118**	**0.417**	**0.050**
3-scale w/o H	**0.866**	0.975	**0.994**	0.120	**0.417**	0.051
3-scale w/o H/8	0.864	0.975	**0.994**	0.121	0.418	0.051
4-scale w/o attention	0.855	0.974	0.993	0.123	0.049	0.052

horizontal flipping and rotation, as well as color adjustment. As for the image size, we crop the image to 352×704 for the KITTI dataset and 416×544 for the NYU V2 dataset.

3.2 Performance Evaluation

The quantitative results of the evaluation on the KITTI dataset are shown in Table 1. It can be seen that the proposed method outperforms other methods on most metrics except for a slight disadvantage on the root mean square error (RMSE) metric. Also, we can see that ResNet 50, ResNeXt 50, and DenseNet 121 are with similar performance, while DenseNet 161 can achieve the best performance due to its bigger capacity.

We also show the quantitative results on the NYU V2 dataset in Table 2. According to the results, the proposed method shows better performance for all metrics except a slight disadvantage on the absolute relative error (AbsRel). Fig. 4 gives some qualitative results. We can see that, HMA-Depth can better understand the relationship among objects (as the walls in the first and second rows), and it can extract better local details (the bookshelf in the third row).

3.3 Ablation Study

To look for the optimal parameters, we conduct an experiment with three variants of the HMA-Depth model. The first two are variants using three scales, rather than four scales, by removing scale H and $H/8$, respectively. In addition, we make another variant by removing all the attention modules to show the significance of hierarchical multi-scale attention, in which the final output is the average of all intermediate predictions. For all variants as well as the base model, we use ResNet 50 as the backbone network and compare their performance on the NYU V2 dataset. The results are shown in Table 3. We can see that the base model achieves the best performance in all the metrics, which demonstrates the effect of multi-scale attention.

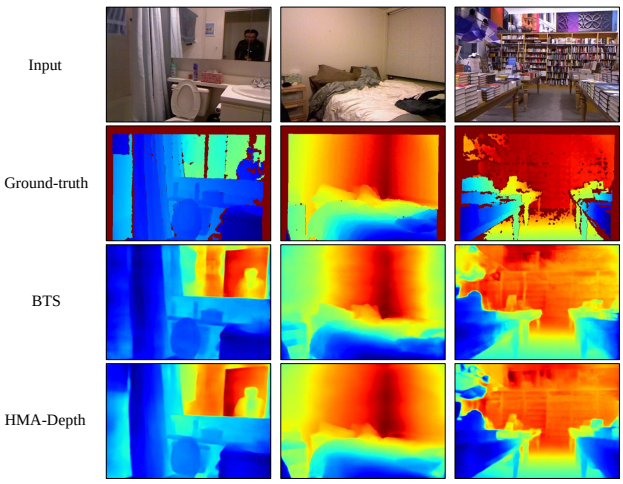

Figure 4. Visualization results of NYU V2 dataset

4 Conclusion

In this paper, we propose a novel network architecture named HMA-Depth that uses a hierarchical multi-scale attention mechanism for monocular depth estimation. For the multi-scale depth maps, attention modules generate the weight masks, indicating which regions in each depth map the model is paying attention to. The experimental results prove the effectiveness of HMA-Depth and show that HMA-Depth outperforms the state-of-the-art methods. However, according to the qualitative results of both KITTI dataset and NYU V2 dataset, we observe that it is not smooth enough on some object surfaces. In the future work, we plan to utilize semantic segmentation results in the depth estimation task, aiming to improve the performance further.

5 Acknowledgment

This research is partially supported by Initiative on Promotion of Supercomputing for Young or Women Researchers, Information Technology Center, The University of Tokyo.

References

[1] Vivek Pradeep, Christoph Rhemann, Shahram Izadi, Christopher Zach, Michael Bleyer, and Steven Bathiche, "MonoFusion: Real-time 3D reconstruction of small scenes with a single web camera," in *IEEE ISMAR*, 2013, pp. 83–88.

[2] Silvan Weder, Johannes Schonberger, Marc Pollefeys, and Martin R Oswald, "RoutedFusion: Learning real-time depth map fusion," in *IEEE CVPR*, 2020, pp. 4887–4897.

[3] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab, "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," in *IEEE CVPR*, 2017, pp. 6243–6252.

[4] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz, "Neural RGB->D sensing: Depth and uncertainty from a video camera," in *IEEE CVPR*, 2019, pp. 10986–10995.

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, "Deep ordinal regression network for monocular depth estimation," in *IEEE CVPR*, 2018, pp. 2002–2011.

[6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng, "Single-image depth perception in the wild," *arXiv preprint arXiv:1604.03901*, 2016.

[7] Rui Wang, Stephen M Pizer, and Jan-Michael Frahm, "Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth," in *IEEE CVPR*, 2019, pp. 5555–5564.

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow, "Digging into self-supervised monocular depth estimation," in *IEEE ICCV*, 2019, pp. 3828–3838.

[9] Arun CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad, "DepthNet: A recurrent neural network architecture for monocular depth prediction," in *IEEE CVPR Workshops*, 2018, pp. 283–291.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE CVPR*, 2017, pp. 270–279.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from RGBD images," in *ECCV*, 2012, pp. 746–760.

[13] 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*, 2014, pp. 2366–2374.

[14] Andrew Tao, Karan Sapra, and Bryan Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[15] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[17] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE CVPR*, 2016, pp. 3640–3649.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[20] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *IEEE CVPR*, 2017, pp. 1492–1500.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 4700–4708.

[22] Ashutosh Saxena, Min Sun, and Andrew Y Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE TPAMI*, vol. 31, no. 5, pp. 824–840, 2008.

[23] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE TPAMI*, vol. 38, no. 10, pp. 2024–2039, 2015.

[24] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *IEEE CVPR*, 2017, pp. 6647–6655.

[25] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *IEEE ICCV*, 2019, pp. 5684–5693.

[26] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille, "Towards unified depth and semantic prediction from a single image," in *IEEE CVPR*, 2015, pp. 2800–2809.

[27] 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 *IEEE CVPR*, 2015, pp. 1119–1127.

[28] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab, "Deeper depth prediction with fully convolutional residual networks," in *IEEE 3DV*, 2016, pp. 239–248.