# LEARNING DEPTH FROM SINGLE IMAGE USING DEPTH-AWARE CONVOLUTION AND STEREO KNOWLEDGE

*Zhenyao Wu*[1], *Xinyi Wu*[1], *Xiaoping Zhang*[2,3], *Song Wang*[1,*], *Lili Ju*[1,*]

[1]University of South Carolina, USA; [2]Wuhan University, China; [3]Farsee2 Technology Ltd, China

{zhenyao, xinyiw}@email.sc.edu, xpzhang.math@whu.edu.cn, songwang@cec.sc.edu, ju@math.sc.edu

## ABSTRACT

Estimating depth from a monocular image has become a very popular task in computer vision for identifying important geometric information of the scene. While its performance has been significantly improved by convolutional neural networks (CNNs) in recent years, depth-estimation accuracy is still unsatisfactory at locations with abrupt depth changes. This is mainly caused by the use of spatially consistent filters in CNNs which directly mix the features of different objects when applied to the pixels near the object borders. Moreover, the performance gap between depth estimation from single image and that from a stereo pair remains quite large due to the ill-posed nature of the former one. In this paper, we propose a new depth-aware convolutional neural network (DACNN) to address these issues. We first design a novel depth-aware convolution operation for DACNN, that can adaptively choose subsets of relevant features for convolutions at each location. Specifically, we compute hierarchical depth features as the guidance, and then estimate the depth map using such depth-aware convolution which can leverage the guidance to adapt the filters. In addition, we also introduce a pre-trained stereo network into DACNN as the teacher to carry out knowledge distillation on the student monocular network with a specially designed loss function. Experimental results on the KITTI online benchmark and Eigen split datasets show that the proposed method achieves the state-of-the-art performance for single-image depth estimation.

***Index Terms***— Depth Estimation, Knowledge Distillation

## 1. INTRODUCTION

Learning depth from a single image is an intriguing computer vision problem and has many important applications. Compared with depth estimation from a stereo pair of images [1] or video sequences [2], inferring accurate depth from single image is much more challenging without the help of multiple view information.

In early years, many approaches make use of Markov Random Fields (MRF), semantic classifiers and superpixels to tackle the single-view depth estimation task. Later, Eigen *et al*. [3] first proposed the use of a multi-scale convolutional architecture to learn depth from single image based on deep learning techniques. Following this innovative work, many more approaches based on convolutional neural networks (CNNs) [4, 5, 6, 7, 8, 9, 10] have been proposed for monocular depth estimation. However, all of these methods treat the features in different depth equally using traditional convolution operations and these convolution operations may mix the features from different objects, which might cause inaccurate prediction of depth and abrupt depth change near the border of two adjacent objects in the image. Inspired by the work of [11] which proposes a segmentation-aware CNN by adapting its filters at each pixel based on segmentation cues, we design a novel depth-aware convolution operation for single-view depth estimation based on depth cues.

Moreover, as an ill-posed problem, single-image depth estimation still shows a very large performance gap from the depth estimation using a pair of stereo images. This is no strange because the former one lacks the crucial multi-view geometric information, even if the use of deep learning techniques can help infer geometric information with data-driven approaches. In this paper, we also propose to make use of the feature extracted from the stereo pair to rectify the ill-posed features extracted from a single image by using the knowledge-distillation technique [12], which was initially proposed for model compression. Previous works [13, 14, 15] have leveraged the distillation to help depth estimation, e.g., Guo *et al*. [13] use pre-trained stereo matching network as a proxy to provide a supervision for the monocular depth estimation. Similarly, Tosi *et al*. [14] use the traditional Semi-Global Matching (SGM) approach to calculate accurate proxy labels for the same purpose. Pilzer *et al*. [15] propose to use the principle of distillation to transfer knowledge from their whole network to the student network which is a part of the teacher network. Different from these existing approaches, the proposed method enforces not only the output similarity, but also the intermediate-feature similarity across the pre-trained stereo network and the student network, with an expectation of further reducing the performance gap between the single-image and the stereo-image depth estimations.

The framework of our depth-aware convolutional neural

*Co-corresponding authors.

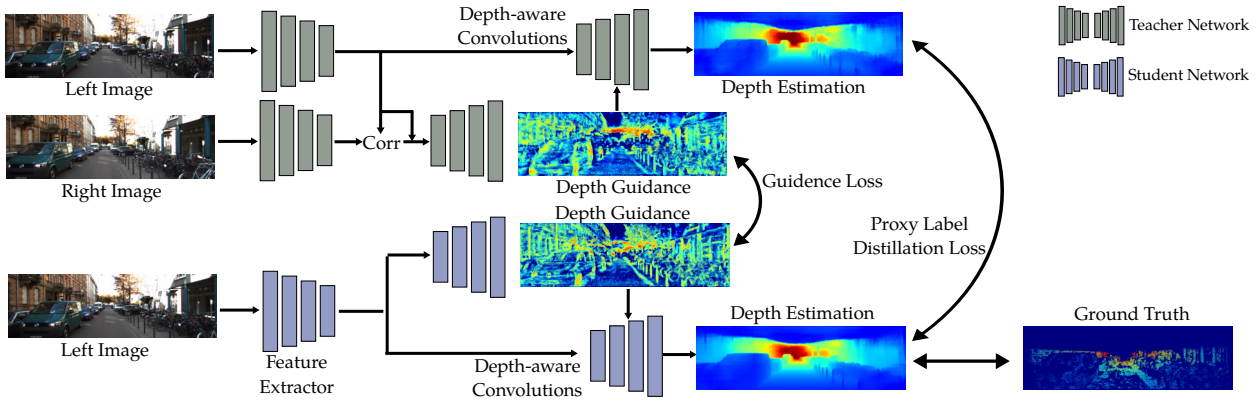


**Fig. 1**. Framework of the proposed DACNN. The pre-trained stereo network (teacher) shown in the top takes the stereo image pair as the input while the monocular network (student) shown at the bottom takes the single image as the input. We constrain both the output similarity and the intermediate-feature similarity across the teacher and the student.

network (DACNN) is illustrated in Figure 1, which consists of a pre-trained stereo network as the teacher and a monocular depth estimation network as a student. Overall the main contributions of our work in this paper include: Firstly, we design a novel depth-aware convolution operation in DACNN to learn the depth with the help of depth cues. Secondly, we also introduce a pre-trained stereo network into DACNN to provide additional supervision on both intermediate features and output of the student through knowledge distillation. Thirdly, our method achieves the state-of-the-art performance for single-image depth estimation on the KITTI online benchmark [16] and the KITTI Eigen split [3].

## 2. RELATED WORK

**Supervised monocular depth estimation** As large-scale datasets (e.g. KITTI [16]) are available, more and more supervised approaches have been developed for monocular depth estimation. Eigen *et al.* [3] propose a multi-scale deep network to learn the depth from a single image by global coarse prediction and local refinement. Follow this paper, several further works have been developed to extract features using deep learning techniques. Fu *et al.* propose DORN [6] by discretizing the depth and converting the regression problem into a multi-class classification problem, which achieves the state-of-the-art performance. Jiao *et al.* [17] propose an attention-driven loss and a synergy network to mutually improve the depth estimation and semantic labeling tasks. Gan *et al.* [7] explicitly model the relationship across different pixels using an additional affinity layer to model the depth relation of neighboring pixels. Recently, Yin *et al.* [9] propose virtual normal directions to incorporate geometric constraints in the 3D space to improve the depth prediction accuracy.

**Knowledge distillation** Knowledge distillation [12] technique is initially proposed for model compression, i.e., transferring knowledge from a cumbersome model to a light-weight model. Later this approach is also taken for knowledge transfer across different domains [18]. Knowledge used for distillation and transfer can be softened labels [12, 19] or intermediate features [20, 21]. Until now, knowledge distillation has been widely applied in computer vision applications, such as object detection [22], pedestrian re-identification [23], and semantic segmentation [24, 25].

## 3. OUR APPROACH

The framework of our method DACNN has been presented in Figure 1, which consists of a pre-trained stereo network (teacher) and a monocular network (student). In this section, we elaborate on the proposed depth-aware convolution operation and the specially designed distillation loss function.

### 3.1. Feature extraction

Both the teacher and student networks use ResNest-50 to extract features from the input, followed by Atrous Spatial Pyramid Pooling (ASPP) [26] with dilation rates $(1, 6, 12, 18)$ to further extract features from multiple receptive fields. Specifically, the stereo network (teacher) takes the stereo image pair as the input and the left and right images share weights during the pre-training. The output of the student network is denoted as $f_s$ and that of the teacher network as $f_l$ and $f_r$.

### 3.2. Depth guidance generation

The depth guidance, also referred to as depth cues, is an intermediate depth feature generated from both the teacher and student networks. For the teacher network, the extracted features ($f_l$ and $f_r$) of the stereo pair are passed into a correlation block, consisting of a correlation layer [27], a $3 \times 3$ convolutional layer and a batch normalization layer, to calculate the matching volume. In parallel, the extracted feature from left image ($f_l$) is also passed into the convolutional layer and the batch normalization layer and the result is concatenated with the matching volume to form the output $O_C$.

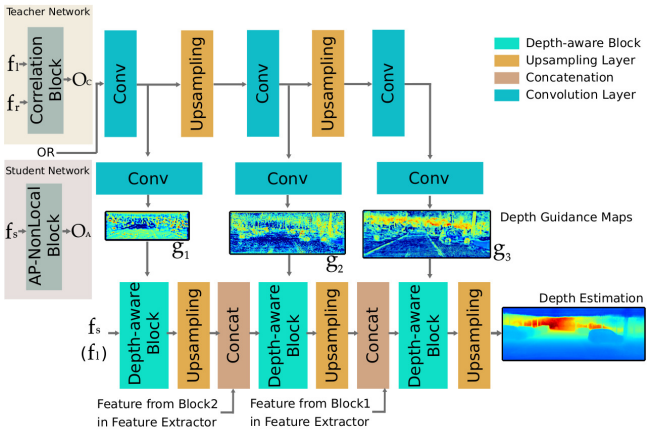


**Fig. 2**. The architecture of the depth guidance estimation branch and the depth map estimation branch in the proposed DACNN. Note that the teacher network and the student network take different input and use different blocks at the beginning of the depth guidance estimation branch. After that, the architectures of the teacher network and the student network are the same.

Here, we consider a maximum displacement of 24 pixels when calculating the matching volume, which corresponds to 192 pixels in the input image. For the student network, we employ the Asymmetric Pyramid Non-Local (AP-NonLocal) Block [28] on the extracted feature of the input single image $f_s$ to obtain the output $O_A$. Then, $O_C$ and $O_A$ are sent to the depth guidance generation branch separately to obtain the depth guidance of the teacher and student networks. The structure of the branch is shown in Figure 2, which contains several $3\times3$ convolution layers and upsampling layers. Finally, we get the depth guidance in three scales for both the teacher and student networks, that are denoted by $g_1^t$, $g_2^t$, $g_3^t$, and $g_1^s$, $g_2^s$ and $g_3^s$, respectively.

### 3.3. Depth-aware convolution

The depth-aware convolution calculates the value for each of the positions based on the depth guidance that was obtained in last step of the previous subsection, and such convolution is employed by both the teacher and student networks.

Let $\mathcal{P}_d = \{-d, 0, d\} \times \{-d, 0, d\}$ represent the receptive field with dilation $d$, then a standard $3 \times 3$ convolutional operation acting on the position $\mathbf{p}$, which takes the single feature $\mathbf{f} \in \mathbb{R}^{C\times H\times W}$ as an input and outputs another feature $\mathbf{f}' \in \mathbb{R}^{C'\times H\times W}$, can be defined by

$$f'_j(\mathbf{p}, d) = \sum_{i=1}^{C} \sum_{\mathbf{o}\in\mathcal{P}_d} w_{i,j}(\mathbf{o}) f_i(\mathbf{p}+\mathbf{o}), \tag{1}$$

where $(w_{i,j}(\mathbf{o})) \in \mathbb{R}^{3\times3\times C\times C'}$ is the weight tensor of a filter, $j = 1, \cdots, C'$, and $C$ and $C'$ are the numbers of channels of the input feature $\mathbf{f}$ and the output feature $\mathbf{f}'$, respectively ( these two features have the same height $H$ and width $W$). Following the previous works [29, 30], we propose the depth-aware convolution across different dilation $d$'s based on depth guidance $g$ as follows:

$$f'_j(\mathbf{p}, d) = \sum_{i=1}^{C} \sum_{\mathbf{o}\in\mathcal{P}_d} w_{i,j}(\mathbf{o}) K(g(\mathbf{p}), g(\mathbf{p}+\mathbf{o})) f_i(\mathbf{p}+\mathbf{o}), \tag{2}$$

where $K$ is a Gaussian operation which makes the convolution to be depth adaptive. For each scale of depth guidance, we use multiple dilations $d$ ($d \in \{1, 6, 12, 18\}$) during the convolution and obtain multi-scale depth features $f'(d)$. These features are concatenated and then fed into a $1 \times 1$ standard convolution layer. All of the above steps compose a depth-aware block. Following this way, we construct multiple depth-aware blocks and upsampling layers to upscale and refine the depth map as shown in Figure 2. Each upsampling layer doubles the resolution of the results and is followed by a $3 \times 3$ standard convolution layer and an ReLU layer. We also concatenate the features from the first two blocks of the feature extractor with the results of upsampling layer to combine the high-level and low-level information. Finally, we obtain the depth map whose resolution is the same as the original image's resolution.

### 3.4. Loss function

To pre-train the teacher network, the loss function is defined by the per-pixel loss $\mathcal{L}_{pixel}$ to measure the distance between the ground truth $d^*_{i,j}$ and the final estimated depth map $d_{i,j}$ , i.e.,

$$\mathcal{L}^T = \mathcal{L}_{pixel} = \frac{1}{N}\sum_{(i,j)}(d_{ij} - d^*_{ij}). \tag{3}$$

For the proposed student network, we also adopt the per-pixel loss $\mathcal{L}_{pixel}$. In addition, we use the smooth loss to encourage the estimated depth map to be locally smooth, which is defined as:

$$\mathcal{L}_{smooth} = \frac{1}{N}\sum_{(i,j)} |\varphi_x d_{ij}| e^{-|\varphi_x I_{ij}|} + |\varphi_y d_{ij}| e^{-|\varphi_y I_{ij}|}, \tag{4}$$

where $I$ is the input image, and the function $\varphi_x$ and $\varphi_y$ calculate the intensity gradients between the neighboring pixels along the $x$ and $y$ directions.

Two more loss functions are proposed for distillation, namely the proxy label transfer loss $\mathcal{L}_{proxy}$ and the guidance transfer loss $\mathcal{L}_{guidance}$, respectively. The former one aims at constraining the output of the student network which uses the estimated result ($\hat{d}_{i,j}$) from the teacher network as a proxy ground truth to coach the student which is defined by

$$\mathcal{L}_{proxy} = \frac{1}{N}\sum_{(i,j)}(d_{ij} - \hat{d}_{ij}). \tag{5}$$

The latter one is to constrain the similarity between the depth guidance from the teacher and the student network. To

**Table 1**. Performance comparison of DACNN and some existing state-of-the-art networks on the KITTI Eigen split.

| Method | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
|---|---|---|---|---|---|---|---|
| Make3D [31] | 0.280 | 3.012 | 8.734 | 0.361 | 0.601 | 0.820 | 0.926 |
| Eigen *et al*. [3] | 0.203 | 1.548 | 6.307 | 0.282 | 0.702 | 0.890 | 0.958 |
| Liu *et al*. [32] | 0.202 | 1.614 | 6.523 | 0.275 | 0.678 | 0.895 | 0.965 |
| Godard *et al*. [33] | 0.114 | 0.898 | 4.935 | 0.206 | 0.861 | 0.949 | 0.976 |
| Kuznietsov *et al*. [34] | 0.113 | 0.741 | 4.621 | 0.189 | 0.862 | 0.960 | 0.986 |
| Gan *et al*. [7] | 0.098 | 0.666 | 3.933 | 0.173 | 0.890 | 0.964 | 0.985 |
| Yin *et al*. [9] | **0.072** | - | 3.256 | 0.117 | 0.938 | **0.993** | **0.998** |
| DORN [6] | **0.072** | 0.307 | **2.727** | 0.120 | 0.932 | 0.984 | 0.994 |
| Our method (DACNN) | 0.073 | **0.304** | 2.801 | **0.116** | **0.939** | 0.990 | 0.997 |

achieve this, a softmax operation is firstly applied to convert the multi-scale guidance maps into distributions, i.e., $p_k^t = \mathrm{softmax}(g_k^t),\ p_k^s = \mathrm{softmax}(g_k^s),\ k = 1, 2, 3$, and then the Kullback-Leibler (KL) divergence is adopted to measure the dissimilarity of the distributions. Specifically, it is defined by

$$\mathcal{L}_{guidance} = \sum_{k=1}^{3} \mathrm{KL}(p_k^t \| p_k^s). \tag{6}$$

The loss function for the student network is finally defined as:

$$\mathcal{L}^S = \mathcal{L}_{pixel} + \alpha \mathcal{L}_{smooth} + \beta \mathcal{L}_{proxy} + \gamma \mathcal{L}_{guidance}, \tag{7}$$

where $\alpha$, $\beta$ and $\gamma$ are weighting factors which are set to 0.01, 0.01 and 1000, respectively.

## 4. EXPERIMENTS

### 4.1. Datasets and evaluation metrics

We use the following popular datasets in experiments for performance evaluation and comparison of the proposed method with many existing state-of-the-art approaches on the monocular depth estimation task.

**KITTI online benchmark** [16]: The KITTI dataset contains over 93K outdoor images and depth maps with the resolution of $1,240 \times 374$. All the images are captured on driving cars by stereo cameras and a Lidar sensor. We use the images from city, residential, road and campus categories to train our model and test on the official test set including 500 images. The scale invariant logarithmic error (SILog), the relative squared error (sqErrorRel), the relative absolute error (absErrorRel) and the root mean squared error of the inverse depth (iRMSE) are used to evaluate the performance on this dataset.

**KITTI Eigen split** [3]: Eigen *et al*. provide a subset of testing split from the KITTI dataset for monocular depth estimation, which is commonly used in recent works. The testing set includes 697 images from 29 scenarios, and we use the images from other 32 scenarios for training. Following [3], we use the absolute relative difference (Abs Rel), the squared relative difference (Sq Rel), the root-mean-square error (RMSE) and the log RMSE (RMSE log) as the error metrics and the accuracy with threshold $\delta = \{1.25, 1.25^2, 1.25^3\}$ as the accuracy metrics. For all error metrics, the lower the better, while for the accuracy metrics, the higher the better.

### 4.2. Experimental settings

The proposed method, DACNN, is implemented using PyTorch, and we pre-train the teacher network and perform the knowledge distillation on the student network using two Nvidia 2080Ti GPUs with the Adam solver (the momentum parameters $\beta_1 = 0.9, \beta_2 = 0.999$). The models are trained from scratch with a batch size of 6. Following [35], we employ the poly learning rate policy from the base learning rate $10^{-4}$ with the power $p = 0.9$. We pre-train the teacher network for 10 epochs and train the student network (distillation) for 15 epochs.

We also perform color normalization on these two datasets for data preprocessing, and during training, all images were randomly cropped to the size of $256 \times 512$. To avoid the overfitting problem, we use the data augmentation strategy in [3]. Specifically, the images are augmented with the random contrast, brightness, and color adjustment in a range of $[0.9, 1.1]$ with 50% of chance.

During the test phase, we split each of the testing image to overlapping windows with the same cropping size as in the training processing, and then obtain the estimated depth values in overlapped regions by averaging the estimations.

### 4.3. Results on the KITTI datasets

The quantitative results evaluated on the KITTI Eigen split are reported in Table 1, which shows that our DACNN achieves the best or close to the best performance in each of the error or accuracy metrics among all compared state-of-the-art networks. To exhibit the visual improvements, we also show some depth estimation results from the test set of Eigen split in Figure 3, from which it is easy to see that the estimated depth maps by our method are much smoother and possess clearer boundary between objects than that by DORN. The quantitative results evaluated from the KITTI online leaderboard are

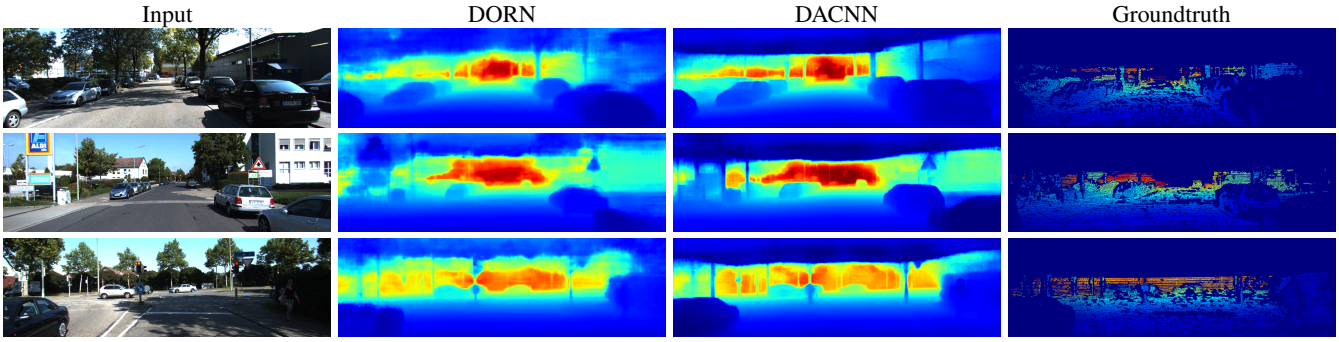


**Fig. 3**. Some depth estimation results from the test set of the KITTI Eigen split. For each row, from left to right are the input image, the estimated depth maps from DORN [6] and our DACNN, and the ground-truth depth map, respectively.

reported in Table 2.

**Table 2**. Performance comparison of DACNN and some existing state-of-the-art networks on the KITTI online benchmark.

| Method | SILog | sqErrorRel | absErrorRel | iRMSE |
|---|---|---|---|---|
| DABC *et al.* [36] | 14.49 | 4.08 | 12.72 | 15.53 |
| Guo *et al.* [13] | 13.41 | 2.86 | 10.60 | 15.06 |
| Zhang *et al.* [37] | 13.08 | 2.72 | 10.27 | 13.95 |
| Yin *et al.* [9] | 12.65 | 2.46 | 10.15 | 13.02 |
| DORN [6] | 11.77 | 2.23 | 8.78 | 12.98 |
| Our method (DACNN) | 12.95 | 2.60 | 10.35 | 13.95 |

### 4.4. Ablation study

In this study, in order to demonstrate effectiveness of the proposed depth-aware blocks in DACNN and the knowledge distillation from the teacher network, we conduct ablation studies to compare some model variants for DACNN on the Eigen split of KITTI dataset. The results are reported in Table 3, from which we can see that the depth-aware blocks are useful for improving the monocular depth estimation, and the knowledge distillation from the teacher network can further improve the overall performance.

**Table 3**. Performance comparison of some model variants of DACNN on the KITTI Eigen split.

| Method | Abs Rel | Sq Rel | RMSE | RMSE log |
|---|---|---|---|---|
| w/o the depth-aware-blocks | 0.086 | 0.563 | 3.675 | 0.162 |
| w/o the knowledge distillation | 0.079 | 0.462 | 3.174 | 0.126 |
| Full version of DACNN | 0.073 | 0.304 | 2.801 | 0.116 |

## 5. CONCLUSION

In this paper, we develope a new depth-aware convolution operation to learn depth from single image by leveraging depth cues. In addition, we incorporate a pre-trained stereo network as a teacher to provide additional supervision for the features and output generated by the student network which is a monocular depth estimation network. Experimental results on the KITTI Eigen split and online benchmark demonstrate that the proposed method can significantly improve the accuracy of monocular depth estimation.

## 6. REFERENCES

[1] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry, "End-to-end learning of geometry and context for deep stereo regression," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 66–75.

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow, "Digging into self-supervised monocular depth prediction," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[3] 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.

[4] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1119–1127.

[5] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, "Deep ordinal regression network for monocular depth estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2002–2011.

[7] Yukang Gan, Xiangyu Xu, Wenxiu Sun, and Liang Lin, "Monocular depth estimation with affinity, vertical pooling, and label enhancement," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 224–239.

[8] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *IEEE Conference on Com-*

*puter Vision and Pattern Recognition (CVPR)*, 2017, pp. 5354–5362.

[9] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5684–5693.

[10] 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.

[11] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos, "Segmentation-aware convolutional networks using local attention masks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5038–5047.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[13] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 484–500.

[14] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9799–9809.

[15] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9768–9777.

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[17] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *European Conference on Computer Vision (ECCV)*, September 2018.

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning*, 2018, pp. 1989–1998.

[19] Nikolaos Passalis and Anastasios Tefas, "Learning deep representations with probabilistic knowledge transfer," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 268–284.

[20] Adriana Romero, Samira Ebrahimi Kahou, Polytechnique Montral, Y. Bengio, Universit De Montral, Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations*, 2015.

[21] Jangho Kim, SeongUk Park, and Nojun Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Advances in Neural Information Processing Systems*, 2018, pp. 2760–2769.

[22] Quanquan Li, Shengying Jin, and Junjie Yan, "Mimicking very efficient network for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6356–6364.

[23] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[24] Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng, "Improving fast segmentation with teacher-student learning," *arXiv preprint arXiv:1810.08476*, 2018.

[25] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang, "Structured knowledge distillation for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2604–2613.

[26] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[27] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2758–2766.

[28] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai, "Asymmetric non-local neural networks for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 593–602.

[29] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz, "Pixel-adaptive convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Anne S. Wannenwetsch and Stefan Roth, "Probabilistic pixel-adaptive refinement networks," in *CVPR*, 2020.

[31] Ashutosh Saxena, Min Sun, and Andrew Y Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

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

[33] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.

[34] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6647–6655.

[35] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang, "Denseaspp for semantic segmentation in street scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3684–3692.

[36] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang, "Deep attention-based classification network for robust depth prediction," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 663–678.

[37] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4106–4115.