End-to-End Disparity Estimation with Multi-granularity Fully Convolutional Network

Guorun Yang and Zhidong Deng(✉)

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science, Tsinghua University, Beijing 100084, China
ygr13@mails.tsinghua.edu.cn, michael@mail.tsinghua.edu.cn

Abstract. Disparity estimation is a challenging task in the field of computer stereo vision. In this paper, we propose a multi-granularity fully convolutional network architecture for end-to-end dense disparity estimation. First, we use single well-pretrained residual network for extraction of multi-granularity and multi-layer features. Second, correlation layers at three different granularities are used to gain hierarchical matching cues between left and right feature maps. Third, we conduct concatenation-deconvolution operations to output disparity maps. Finally, the experimental results show that our method achieves state of the art results, taking the second place on the KITTI Stereo 2012 task.

Keywords: Multi-granularity · Correlation · Concatenation-deconvolution · Disparity estimation

1 Introduction

Disparity estimation is a classical problem in the field of stereo vision. It has been extensively applied to many areas such as view synthesis, object detection, and robot navigation. The main goal of disparity estimation is to calculate the displacement of corresponding pixels between left and right images, where corresponding pixels result from identical 3D points projected onto the two image planes. Displacement values at each location forms so-called disparity map.

It is challenging to perform disparity estimation accurately, particularly predict dense disparity map. The majority of stereo algorithms treat such task as a matching problem, measuring similarity between two corresponding patches of left and right images. From this point of view, the main idea of those algorithms is to develop powerful feature representation for image patches. Then the resulting feature vectors can be employed to compute match cost and then pick the best matching pixel between left and right images. In recent years, deep convolutional neural network (CNN) has demonstrated remarkable performance in many fields including computer vision, speech recognition, natural language processing, self-driving, and big data analysis through representation learning of

D. Liu et al. (Eds.): ICONIP 2017, Part III, LNCS 10636, pp. 238–248, 2017.
https://doi.org/10.1007/978-3-319-70090-8_25

hierarchical features based on a large scale labeled data. With utilization of CNN features, such patch-based matching methods can be significantly improved in terms of accuracy of disparity prediction.

Except for disparity calculation based on similarity of image patches, dense disparity map estimation problem could be considered as a pixel-wise labeling task, where each pixel would be assigned a real-value disparity. Lately, inspired by the success of fully-convolutional network (FCN) [1] for semantic segmentation task, such an end-to-end learning structure was introduced to predict disparity map [2]. The combination of encoder (top-down) and decoder (bottom-up) architecture can effectively link the global scene information with local disparity estimation, which leads to further improvements in both accuracy and speed.

In general, FCN models for disparity estimation contain a correlation module to extract matching information from left and right feature maps. Several approaches like [2,3] deploy correlation operations on low-level feature maps. In our opinion, matching cues not only exist in low-level features, but also occur in high-level features. Furthermore, the category information in high-level feature maps could be utilized to compensate matching cues lost in low-level features. For example, in an urban scene, adjacent road and sidewalk are difficult to distinguish from low-level features due to similar colors and textures. However, it would be convenient to differentiate in high-level semantic features. As a result, this paper attempts to extract matching cues from multiple granularity feature maps and aggregate them together. In the proposed method, we first exploit well-pretrained ResNet-50 [4] to obtain different granular hierarchical features. Second, different granular correlation layers are presented to produce feature maps that embed a diversity of matching cues. Finally, we design a concatenation-deconvolution sub-structure to aggregate all the matching information from different granularities and carry out regression of pixel-level disparity values.

The main contributions of this paper are summarized below:

- We learn to represent three different granularities of matching information.
- Those matching cues are aggregated to enhance capabilities of stereo disparity regression.
- On the KITTI Stereo 2012 task [5], the proposed multi-granularity FCN achieves state-of-the-art performance.

2 Related Work

There has been a large amount of work on stereo disparity estimation. In [6] proposed by Scharstein et al., stereo algorithms are regarded to generally include the following four steps: matching cost computation, cost aggregation, disparity computation, and disparity refinement. Several local descriptors based on gradient or binary patterns are designed to compute local matching cost [7,8], accompanying by some global optimization methods to improve results [9].

Zbontar and LeCun [10] used CNN for matching cost computation. Luo et al. [11] proposed a siamese network that extracted marginal distributions

over all possible disparities for each pixel. Chen et al. [12] presented a multi-scale deep embedding model that fused features vectors learned within different scale-spaces. Shaked and Wolf [13] proposed a highway network architecture with a hybrid loss that conducted multi-level comparison of image patches.

Inspired by other pixel-wise labeling tasks such as semantic segmentation [1,14,15], the FCN is introduced for the end-to-end learning of disparity map. In 2016, Mayer et al. [2] proposed DispNet for real time disparity estimation. There is a structure similar to their previous work called FlowNet [3], which directly inspires us to use correlation layers for encoding matching cues.

Lately, several researchers extended FCN architecture to make further improvement for disparity or depth estimation. Kendall et al. [16] proposed an architecture called GC-Net that incorporates contextual information by means of 3D convolutions over a cost volume. Gidaris and Komodakis [17] presented a cascade network that had a pipeline to detect, replace, and refine the predicted errors. Kuznietsov et al. [18] proposed a semi-supervised approach for monocular depth map prediction. During training phase, they not only use ground-truth depth for supervised learning, but also define an alignment loss based on photo consistency. In this paper, we first employ well-pretrained ResNet-50 to have extraction of multi-scale and multi-layer feature maps and then adopt three different granularitie of correlation layers to get a diversity of matching information. Those matching cues are further aggregated to improve performance of stereo disparity regression. Finally, on the KITTI Stereo 2012 task [5], the proposed multi-granularity FCN achieves state-of-the-art results, ranking second compared to the other 94 competitors.

3 Model Architecture

Our multi-granularity FCN (MG-FCN) architecture is shown in Fig. 1. This is a data-driven model that enables end-to-end disparity learning. It is observed from Fig. 1 that it could be divided into three sub-structures: representation of multi-granularity features, correlation layers and concatenation-deconvolution.

3.1 Representation of Multi-granularity Features

Unlike computing matching cost on the pair of original images, this paper extracts three granularities of hierarchical features of left and right raw images. For this sake, we exploit a ResNet-50 that was well pretrained on a large scale benchmark of ImageNet. ResNet [4] is currently believed as one of the best CNN model due to allowing the network to have much deeper layers. As shown in Fig. 1, such a ResNet-50 model comprises three blocks that output three different granularities of features, respectively, which implies that the feature maps of ${M_1}^L$, ${M_1}^R$, ${M_2}^L$, ${M_2}^R$, ${M_3}^L$, ${M_3}^R$ will be used as inputs of incoming correlation layers.

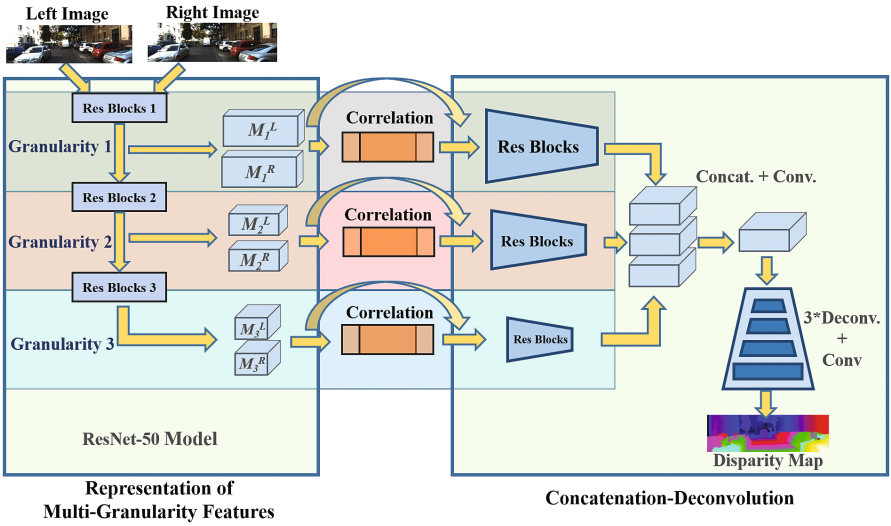


Fig. 1. Our MG-FCN architecture. *Res Blocks* indicate components in residual network, which comprises convolutional, batch normalization, and ReLU layers with split-transform-merge strategy, and the blue cubes represent feature maps. (Color figure online)

3.2 Correlation Layers

The three granularities of correlation layers, which involves the description of matching cost between corresponding patches, are critical in the MG-FCN architecture. Fischer et al. [3] defined a correlation layer in the FlowNet for optical flow estimation. This paper presents three different granularities of correlation layers for a diversity of matching cues. Given one displacement value, correlation layers are used to convolve left and right feature maps of ${M_i}^L$, ${M_i}^R$ $(i = 1, 2, 3)$ and further make summation of resulting multi-channel maps to generate one final matching feature map. The correlation of two patches centered at x_1 in M_i^L and x_2 in M_i^R is defined as

$$c(x_1, x_2) = \sum_{o \epsilon [-k,k] \times [-k,k]} \langle {M_i}^L(x_1 + o), {M_i}^R(x_2 + o) \rangle \tag{1}$$

where $K = 2k + 1$ is the size of patch. We set the maximum displacement β_i $(i = 1, 2, 3)$ to restrict search of possible patch-pairs. The correlation $c(x_1, x_2)$ is only calculated in the neighborhood of size $s_i = \beta_i + 1$, which implies uni-direction searching on ${M_i}^R$. Finally, the size of resulting feature maps for each of three granularities of correlation layers is $(s_i \times w \times h)$, where w indicates the width and h the height.

Table 1. The layers in our Concatenation-deconvolution sub-structure, where **Ch. I/O** denotes channels of input and output feature maps, **Scale** means the scaling factor of output feature map size. The **corr**, **conv**, **concat deconv** and **res** layer denote correlation, convolutional, concatenate, deconvolutional layer and residual blocks respectively. The superscript and subscript of layer indicate the stride and kernel size of convolutional or deconvolutional layer.

Granularity #1			Granularity #2			Granularity #3		
Layer	Ch. I/O	Scale	Layer	Ch. I/O	Scale	Layer	Ch. I/O	Scale
corr_a	128/97	1/2	corr_b	256/49	1/4	corr_c	1024/25	1/8
concat_a	(128+97)/225	1/2	concat_b	(256+49)/305	1/4	concat_c	(1024+25)/1049	1/8
poola_2 ^{3}	225/225	1/4	res_$1b_1$ ^{3}	305/256	1/4	res_$1c_1$ ^{3}	1049/1024	1/8
res_$1a_1$ ^{3}	225/256	1/4	res_$2b_2$ ^{3}	256/512	1/8	res_$2c_1$ ^{3}	1024/1024	1/8
res_$2a_1$ ^{3}	256/256	1/4	res_$3b_1$ ^{3}	512/512	1/8	res_$3c_1$ ^{3}	1024/1024	1/8
res_$3a_1$ ^{3}	256/256	1/4	res_$4b_1$ ^{3}	512/512	1/8	res_$4c_1$ ^{3}	1024/1024	1/8
res_$4a_2$ ^{3}	256/512	1/8	res_$5b_1$ ^{3}	512/512	1/8	res_$5c_1$ ^{3}	1024/1024	1/8
res_$5a_1$ ^{3}	512/512	1/8	res_$6b_1$ ^{3}	512/1024	1/8	res_$6c_1$ ^{3}	1024/2048	1/8
res_$6a_1$ ^{3}	512/512	1/8	res_$7b_1$ ^{3}	1024/1024	1/8	res_$7c_1$ ^{3}	2048/2048	1/8
res_$7a_1$ ^{3}	512/512	1/8	res_$8b_1$ ^{3}	1024/1024	1/8	res_$8c_1$ ^{3}	2048/2048	1/8
res_$8a_1$ ^{3}	512/1024	1/8	res_$9b_1$ ^{3}	1024/1024	1/8	conv_c_1 ^{3}	2048/512	1/8
res_$9a_1$ ^{3}	1024/1024	1/8	res_$10b_1$ ^{3}	1024/1024	1/8			
res_$10a_1$ ^{3}	1024/1024	1/8	res_$11b_1$ ^{3}	1024/1024	1/8			
res_$11a_1$ ^{3}	1024/1024	1/8	res_$12b_1$ ^{3}	1024/2048	1/8			
res_$12a_1$ ^{3}	1024/1024	1/8	res_$13b_1$ ^{3}	2048/2048	1/8			
res_$13a_1$ ^{3}	1024/1024	1/8	res_$14b_1$ ^{3}	2048/2048	1/8			
res_$14a_1$ ^{3}	1024/2048	1/8	conv_b_1 ^{3}	2048/512	1/8			
res_$15a_1$ ^{3}	2048/2048	1/8						
res_$16a_1$ ^{3}	2048/2048	1/8						
conv_a_1 ^{3}	2048/512	1/8						

Layer	Channels I/O	Scale	Inputs
concat	(512+512+512)/1536	1/8	conv_a, conv_b, conv_c
conv_2_1 ^{1}	1536/512	1/8	concat
deconv_1_1 ^{3}	512/256	1/4	conv_2
deconv_2_1 ^{3}	256/128	1/2	deconv_1
deconv_3_1 ^{3}	128/64	1	deconv_2
conv_3_1 ^{1}	64/1	1	deconv_3

3.3 Concatenation-Deconvolution

The concatenation-deconvolution sub-structure is designed to conduct feature aggregation and regression of stereo disparity values based on preceding feature maps that contain three different granularities of matching information from the correlation layers. As shown in Fig. 1 and Table 1, three residual blocks are used to further encode corresponding matching features before concatenation. In order to reduce the number of feature channels, we employ one $1*1$ convolutional layer to merge the concatenated feature maps. Finally, three deconvolutional layers and an extra convolutional layers are adopted to generate stereo deparity values.

The last convolutional layer outputs the predicted disparity maps. For end-to-end learning, it is required to define a loss function to measure the errors between the predicted disparity maps and the ground truths. This paper directly computes the absolute errors (L1-norm) between the predicted values d_i and the

ground-truths $\hat{d}_i$ for each valid disparity pixels. Compared to other norms used for loss functions, we believe that the L1-norm function is more intuitive to describe the deviation between predicted disparities and the ground truths.

$$Loss(I_l, I_r, D) = \frac{1}{N_{\Omega_D}} \sum_{i \epsilon \Omega_D} ||d_i - \hat{d}_i||_1 \tag{2}$$

where Ω_D denotes the set of valid pixels that have the ground truths, N_{Ω_D} the number of valid pixels, I_l the left image, I_r the right image and D the ground truth of disparity map.

4 Experimental Results

We evaluated our method on CityScapes [19] and KITTI Stereo 2012 [5] datasets. Both of the two datasets provide stereo images and disparit ground truths. On CityScapes benchmark, the disparity maps are pre-computed by the SGM algorithm [9]. We use the "$gtFine$" subset that contains 5,000 images. The official split on this subset is that 2,975 images are exploited for training and 500 images are used for validation. On KITTI Stereo 2012 dataset, there are 194 training images with sparse disparity ground truth and 195 test images. To facilitate the comparison among different architectures, we split the training dataset like that of Luo et al. [11], in which 160 images are randomly selected for training and the remaining 34 images are adopted for validation.

In order to verify the performance of the MG-FCN, we compare it with SG-FCNs on both CityScapes and KITTI Stereo 2012. In each of three SG-FCNs, a single-granularity feature is extracted, followed by one correlation layer, residual blocks, and three deconvolution layers to learn disparity.

4.1 Implementation Details

We initially pre-trained the ResNet-50 [4] on a large scale ImageNet dataset. The three feature maps at $conv1$, $pool1$ and $res4a$ layer of well-pretrained ResNet-50 are used for the three granularities of correlation operations. In three granularities of correlation layers, we set the maximum displacement $d = 96, 48, 24$, respectively. In the concatenation-deconvolution sub-structure, we adopted the same initialization procedure as He et al. [20] in the convolution and deconvolution layers involved. Meanwhile, we used Caffe framework and stochastic gradient descent (SGD) with momentum of 0.9 to train the MG-FCN. To avoid overfittings, we employed L2 regularization on the weights with decay of $w_d = 0.0001$.

Considering that the KITTI dataset only contains small and sparse labeled samples, we first trained the MG-FCN on CityScapes dataset with the initial learning rate $lr = 0.01$ and then fine-tuned it on KITTI dataset with the initial

learning rate $lr = 0.001$. We exploited the polynomial learning rate policy with 90k iterations. Moreover, we took a random resize factor of $\alpha \in [0.5, 2.0]$ and the crop size of $513 * 321$ for data augmentation.

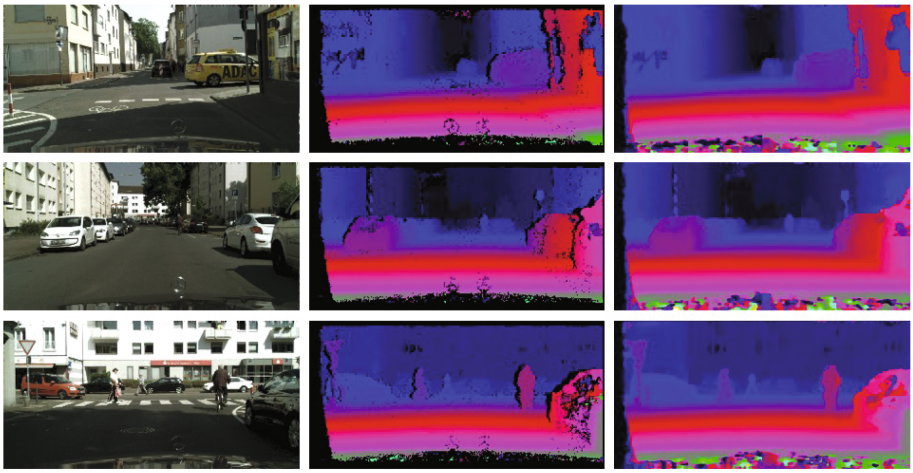
4.2 Results

The experimental results in Table 2 show the test error of three single-granularity FCN (SG-FCN) models and MG-FCN model on the validation dataset of CityScapes and KITTI Stereo. The SG-FCN #1 means that we concatenate correlation layer on *conv*1 feature maps of ResNet-50. The SG-FCN #2 and SG-FCN #3 are linked to *pool*1 and *conv*4*a* feature maps, respectively. The items $> i$ pixels $(i = 1, 2, 3, 4)$ indicate different thresholds adopted to decide whether an estimated disparity value is correct. Numerical results in Table 2 measure the proportion of mistaken disparity pixels. The above comparison demonstrates that the MG-FCN model performs significantly better than the three single-granularity FCN (SG-FCN) models through the aggregation of matching cues on multiple granularities. Figure 2 shows the qualitative results on CityScapes, KITTI validation and test datasets respectively.

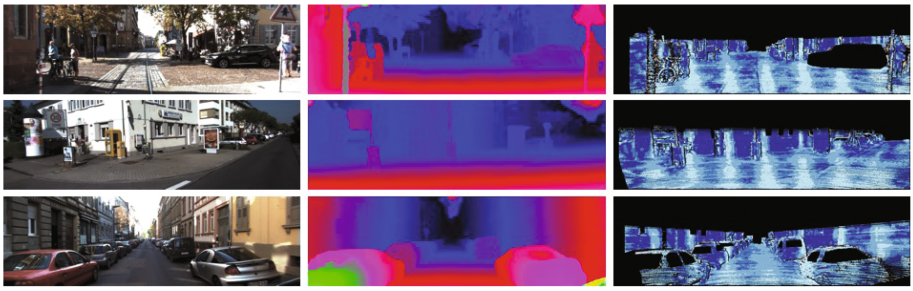
Table 2. The test error of SG-FCNs and MG-FCN across different error thresholds on the CityScapes and KITTI 2012 dataset

	CityScapes				KITTI			
	>2 px	>3 px	>4 px	>5 px	>2 px	>3 px	>4 px	>5 px
SG-FCN#1	5.38	3.16	2.28	1.81	4.61	2.86	2.06	1.60
SG-FCN#2	5.92	3.33	2.37	1.88	5.24	3.12	2.23	1.73
SG-FCN#3	7.20	4.01	2.75	2.12	8.64	4.67	2.89	2.00
MG-FCN	**4.35**	**2.60**	**1.90**	**1.55**	**4.31**	**2.67**	**1.94**	**1.52**

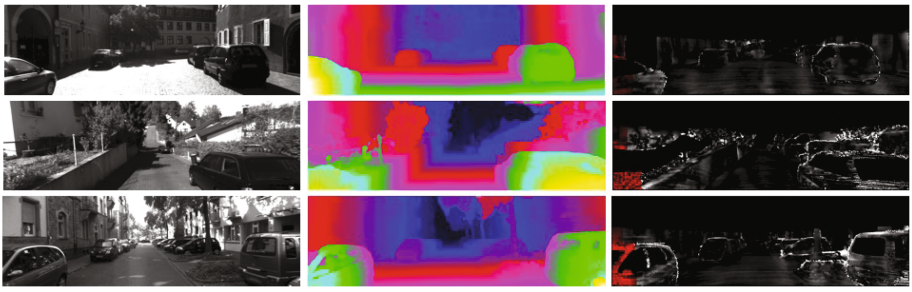
In Table 3, we evaluated our method on KITTI 2012 benchmark [5]. The item "Noc" refers to evaluation on non-occluded regions, i.e., regions for which the matching correspondence is inside the image domain, while "All" refers to evaluation on all image regions for which ground truth could be measured. "End-Point" denote the average end-point deviation between predicted disparity values and ground truth. Our MG-FCN achieves state-of-the-art results, which outperforms most patch-based methods [10,11,21] on both accuracy and runtime. Among FCN methods [2], our model is also competitive, just behind the GC-Net [16], ranking second on the ratings (http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo).



(a) CityScapes data qualitative results on validation dataset. From left: left stereo input image, ground truth, disparity prediction



(b) KITTI data qualitative results on validation dataset. From left: left stereo input image, disparity prediction, error map.



(c) KITTI data qualitative results on test dataset. From left: left stereo input image, disparity prediction, error map.

Fig. 2. Qualitative results. By learning to aggregate multi-granularity matching cues, our method could perform accurate disparity estimation on challenging scenarios.

Table 3. Comparision to state-of-art results on the KITTI 2012 benchmark

	>2 pixels		>3 pixels		>4 pixels		>5 pixels		End-Point		Runtime (s)
	Noc	All	Noc	All	Noc	All	Noc	All	Noc	All	
GC-NET [16]	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46	0.6 px	0.7 px	0.9
L-ResMatch [13]	3.64	5.06	2.27	3.40	1.76	2.67	1.50	2.26	0.7 px	1.0 px	48
PBCP [21]	3.62	5.01	2.36	3.45	1.88	2.74	1.62	2.32	0.7 px	0.9 px	68
Displets v2 [22]	3.43	4.46	2.37	3.09	1.97	2.52	1.72	2.17	0.7 px	0.8 px	265
MC-CNN-arct [10]	3.90	5.45	2.43	3.63	1.90	2.85	1.64	2.39	0.7 px	0.9 px	67
Content-CNN [11]	4.98	6.51	3.07	4.29	2.39	3.36	2.03	2.82	0.8 px	1.0 px	0.7
Deep Embed [12]	5.05	6.47	3.10	4.24	1.73	2.32	1.92	2.68	0.9 px	1.1 px	3
DispNetC [2]	7.38	8.11	4.11	4.65	2.77	3.30	2.05	2.39	0.9 px	1.0 px	0.06
MG-FCN (Ours)	**3.73**	**4.41**	**2.17**	**2.68**	**1.56**	**1.97**	**1.22**	**1.56**	**0.8 px**	**0.8 px**	**0.6**

5 Conclusions

In this paper, we propose a MG-FCN model for end-to-end disparity estimation. In such a new pixel-level disparity prediction method, one ResNet-50 that was well pretrained on ImageNet is first employed to represent multi-scale and multi-layer features of raw left and right images. Second, we present three different granularities of correlation layers to seek a diversity of matching cues. Thrid, the feature maps that include matching information are concatenated and merged so as to perform stereo disparity regression. We evaluate the performance of the proposed MG-FCN model on both CityScapes and KITTI Stereo 2012 dataset. Finally, our method achieves state-of-the-art results on KITTI Stereo 2012 benchmark. In the future, we will focus on semi-supervised or even unsupervised learning methods for such a challenging problem. Meanwhile, it is very interesting to us to compress the above-mentioned model for real-time applications such as self-driving car.

Acknowledgments. This work was supported in part by the National Science Foundation of China (NSFC) under Grant Nos. 91420106, 90820305, and 60775040, and by the National High-Tech R&D Program of China under Grant No. 2012AA041402. We would like to thank Zeping Li and Shiyao Wang for their helps during preparation of this paper.

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 (2015)
2. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) arXiv:1512.02134 (2016)

3. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: FlowNet: learning optical flow with convolutional networks. In: IEEE International Conference on Computer Vision (ICCV) (2015)
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 (2016)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
6. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: Proceedings of IEEE Workshop on Stereo and Multi-Baseline Vision, 2001, (SMBV 2001), pp. 131–140. IEEE (2001)
7. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6492, pp. 25–38. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19315-6_3
8. Heise, P., Jensen, B., Klose, S., Knoll, A.: Fast dense stereo correspondences by binary locality sensitive hashing. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 105–110. IEEE (2015)
9. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008)
10. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. **17**, 1–32 (2016)
11. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5695–5703 (2016)
12. Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C.: A deep visual correspondence embedding model for stereo matching costs. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 972–980 (2015)
13. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. arXiv preprint arXiv:1701.00165 (2016)
14. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
15. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly- and semi-supervised learning of a DCNN for semantic image segmentation arXiv:1502.02734 (2015)
16. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. arXiv preprint arXiv:1703.04309 (2017)
17. Gidaris, S., Komodakis, N.: Detect, replace, refine: deep structured prediction for pixel wise labeling. arXiv preprint arXiv:1612.04770 (2016)
18. Kuznietsov, Y., Stückler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. arXiv preprint arXiv:1702.02706 (2017)
19. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
20. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)

21. Seki, A., Pollefeys, M.: Patch based confidence prediction for dense disparity map. In: British Machine Vision Conference (BMVC), vol. 10 (2016)
22. Guney, F., Geiger, A.: Displets: resolving stereo ambiguities using object knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4165–4175 (2015)