Supplementary material for StereoDRNet

1. Introduction

In this supplementary material, we provide additional details of the training and evaluation procedure of our indoor scene reconstruction experiments. We also provide in depth detail of our proposed network architecture and show the effect of the proposed refinement procedure on the reconstruction quality. We share the results of the ablation study on the dilated convolutions used in our cost filtering approach and visualize the comparison of the disparity predictions from our system with state of art methods on KITTI and ETH3D benchmarks.

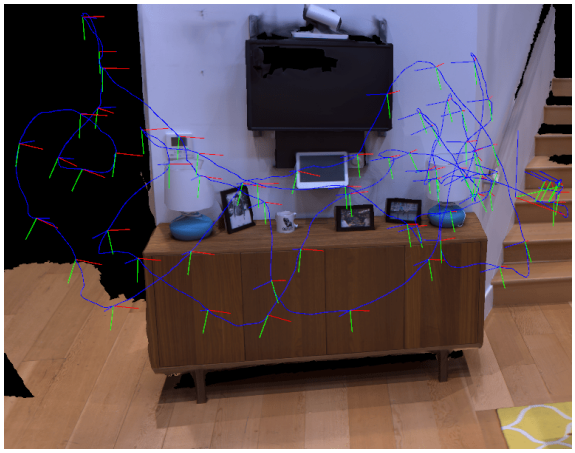

Figure 1: This figure shows the training scene used for all our indoor scene reconstruction experiments. We used about 200 stereo views rendered by OpenGL from poses along the camera trajectory visualized by the blue curve. The 3D reconstruction was built using the method described in [10].

2. 3D Reconstruction Experiments

For all 3D reconstruction experiments and evaluations we used a set of about 200 stereo views shown in Fig. 1 to fine tune the SceneFlow [5]-pre-trained networks.

We show the textured 3D reconstructions of our indoor scene dataset in Fig. 2. Note that we used KinectFusion [6] to fuse the depth maps into 3D spatial maps. We did not use any structure-from-motion (SfM) or external localization method for estimating camera trajectories. Hence, the camera views visualized in Fig. 2 are the output of the ICP (iterative closest point) procedure used by the KinectFusion [6] system. We used manual adjustment followed by ICP to align the 3D reconstructions wherever necessary for our evaluations.

3. Network Details

We provide the network architecture of StereoDRNet in Table. 3. We borrowed ideas on extracting robust local image features from PSMNet [1]. As described in the paper, we use Vortex Pooling [11] for extracting global scene context. In our experiments we found dilation rates 3, 5 and 15 and average grids of size 3×3, 5×5 and 15×15 to improve performance more in disparity predictions than the one proposed in the original work for semantic segmentation.

3D Dilation in Cost Filtering				SceneFlow
rate = 1	rate = 2	rate = 4	rate = 8	EPE
✓				1.13
✓	✓			1.03
✓	✓	✓		**0.98**
✓	✓	✓	✓	1.01

Table 1: Ablation study of dilated convolution rates used in the proposed dilated cost filtering scheme. Note that we used StereoDRNet without refinement in this study.

In order to show the effectiveness of the proposed dilated convolutions in cost filtering, we conduct an ablation study in Table. 1 on the SceneFlow [5] dataset. We observed that increasing dilation rates improved the quality of predictions. Dilation rates above 4 did not provide any significant gains.

The proposed refinement network described in Table. 2 is inspired by the refinement procedures proposed in CRL [7], iResNet [4], StereoNet [3], and ActiveStereoNet [13]. We adopted the basic architecture for refinement as described in StereoNet [3] with dilated residual blocks [12] to increase the receptive field of filtering without compromising resolution. This technique was also adopted in recent work on optical flow prediction Pwc-net [9]. We experienced additional gains when using the photometric error E_p and geometric error maps E_g as in-

Index	Layer Description	Output
1	Warp(I_R,$\mathbf{d_L^3}$) - I_L	H x W x 3
2	concat 1, I_L	H x W x 6
3	Warp($\mathbf{d_R^3}$, $\mathbf{d_L^3}$) - $\mathbf{d_L^3}$	H x W x 1
4	concat 3, $\mathbf{d_L^3}$	H x W x 2
5	3x3 conv on 2, 16 features	H x W x 16
6	3x3 conv on 4, 16 features	H x W x 16
7	concat 5,6 I_L	H x W x 32
8-13	(3x3 conv, residual block) x 6, dil rate 1,2,4,8,1,1	H x W x 32
14	3x3 conv, 2 features as 14(a) and 14(b)	H x W x 2
15	$\mathbf{d^r}$: 14(a) + $\mathbf{d_L^3}$	H x W
16	$\mathbf{O}$: sigmoid on 14(b)	H x W

Table 2: Refinement network for StereoDRNet. $\mathbf{d^r}$ and $\mathbf{O}$ represent refined disparity and occlusion probability respectively.

puts and co-training of occlusion maps. Such enhancements in the refinement procedure has never been proposed to the best of our knowledge.

4. Effect of Refinement

Our refinement procedure not only improves the overall disparity error but also makes the prediction geometrically consistent. We calculate surface normal maps from disparity/depth maps using the approach described in KinectFusion [6]. We use a surface normal error metric to measure consistency in the disparity predictions (first order derivative). Figures 3 and 4 visualize how our refinement procedure improves the overall structure of objects. In some cases such as in the first comparison in Fig. 3 we observe little improvement in disparity prediction but large improvement in surface normals. Figure 4 demonstrates real scene disparity and derived surface normal predictions and proves that our refinement procedure works well on real world data in presence of shadows and dark lighting conditions. Dense 3D reconstruction methods such as KinectFusion [6] use surface normals to calculate fusion parameters and confidence weights, hence it is important to predict geometrically consistent disparity or normal maps for high quality 3D reconstruction.

Index	Layer Description	Output
1	Input Image	H x W x 3
	Local feature extraction	
2	3x3 conv, 32 features, stride 2	H/2 x W/2 x 32
3-4	(3x3 conv, 32 features) x 2	H/2 x W/2 x 32
5-7	(3x3 conv, 32 features, res block) x 3	H/2 x W/2 x 32
8	3x3 conv, 32 features, stride 2	H/4 x W/4 x 32
9-22	(3x3 conv, 64 features, res block) x 15	H/4 x W/4 x 64
23-28	(3x3 conv, 128 features, res block) x 6	H/4 x W/4 x 128
	Spatial Pooling	
29	Global Avg Pool on 28, bi-linear interp	H/4 x W/4 x 128
30	Avg Pool 3x3 on 28, conv 3x3, dil rate 3	H/4 x W/4 x 128
31	Avg Pool 5x5 on 28, conv 3x3, dil rate 5	H/4 x W/4 x 128
32	Avg Pool 15x15 on 28, conv 3x3, dil rate 15	H/4 x W/4 x 128
33	Concat 22, 28, 29, 30, 31 and 32	H/4 x W/4 x 704
34	3x3 conv, 128 features	H/4 x W/4 x 128
35	1 x 1 conv, 32 features without BN and ReLU	H/4 x W/4 x 32
	Cost Volume	
36	Subtract left 35 from right 35 with D/4 shifts,vice versa	D/4 x H/4 x W/4 x 64
	Cost Filtering	
37-38	(3x3x3 conv, 32 features) x 2	D/4 x H/4 x W/4 x 32
39	3x3x3 conv, 32 features, stride 2	D/8 x H/8 x W/8 x 32
40	3x3x3 conv, 32 features	D/8 x H/8 x W/8 x 32
41	3x3x3 conv on 39, 32 features	D/8 x H/8 x W/8 x 32
42	3x3x3 conv on 39, 32 features, dil rate 2	D/8 x H/8 x W/8 x 32
43	3x3x3 conv on 39, 32 features, dil rate 4	D/8 x H/8 x W/8 x 32
44	3x3x3 conv on concat(41,42,43), 32 features	D/8 x H/8 x W/8 x 32
45	3x3x3 deconv, 32 features, stride 2	D/4 x H/4 x W/4 x 32
46	**Pred1**: 3x3x3 conv on 45 + 38	D/4 x H/4 x W/4 x 2
47	3x3x3 conv on 45, 32 features, stride 2	D/8 x H/8 x W/8 x 32
48	3x3x3 conv + 40, 32 features	D/8 x H/8 x W/8 x 32
49	3x3x3 conv on 48, 32 features	D/8 x H/8 x W/8 x 32
50	3x3x3 conv on 48, 32 features, dil rate 2	D/8 x H/8 x W/8 x 32
51	3x3x3 conv on 48, 32 features, dil rate 4	D/8 x H/8 x W/8 x 32
52	3x3x3 conv on concat(49,50,51), 32 features	D/8 x H/8 x W/8 x 32
53	3x3x3 deconv, 32 features, stride 2	D/4 x H/4 x W/4 x 32
54	**Pred2**: 3x3x3 conv on 53 + 38	D/4 x H/4 x W/4 x 2
55	3x3x3 conv on 53, 32 features, stride 2	D/8 x H/8 x W/8 x 32
56	3x3x3 conv + 48, 32 features	D/8 x H/8 x W/8 x 32
57	3x3x3 conv on 56, 32 features	D/8 x H/8 x W/8 x 32
58	3x3x3 conv on 56, 32 features, dil rate 2	D/8 x H/8 x W/8 x 32
59	3x3x3 conv on 56, 32 features, dil rate 4	D/8 x H/8 x W/8 x 32
60	3x3x3 conv on concat(57,58,59), 32 features	D/8 x H/8 x W/8 x 32
61	3x3x3 deconv, 32 features, stride 2	D/4 x H/4 x W/4 x 32
62	**Pred3**: 3x3x3 conv on 61 + 38	D/4 x H/4 x W/4 x 2
	Disparity Regression	
63	Bi-linear interp of **Pred1**, **Pred2**, **Pred3**	D x H x W x 2
64	SoftArg Max of 63 to get $\mathbf{d^1}$, $\mathbf{d^2}$, $\mathbf{d^3}$	H x W x 2

Table 3: Full StereoDRNet architecture. Note that when used without refinement, StereoDRNet just outputs $\mathbf{d^1}$, $\mathbf{d^2}$ and $\mathbf{d^3}$ for the left view.

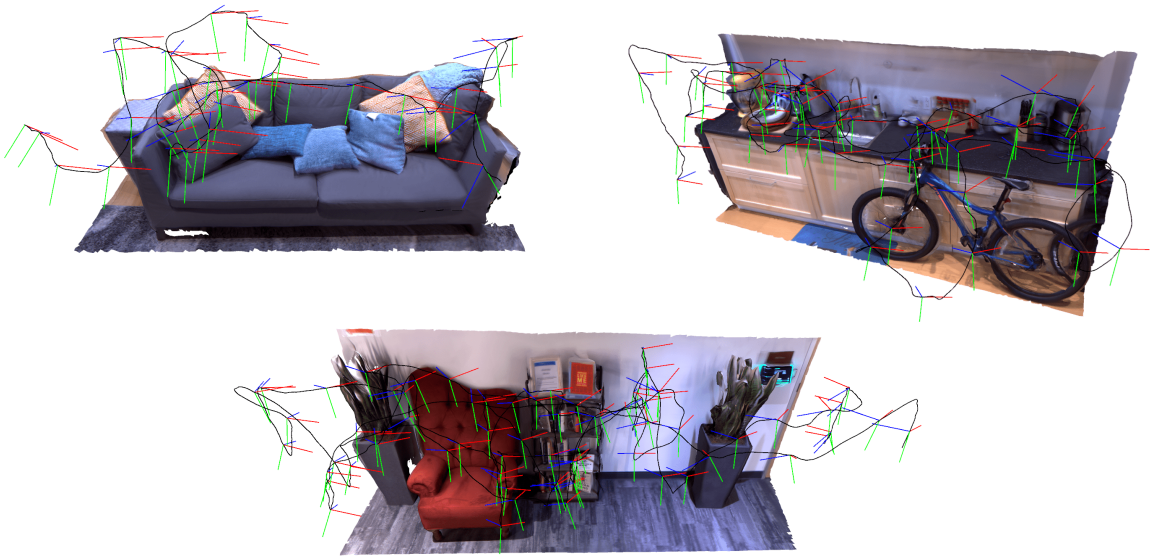

Figure 2: This figure shows the textured 3D reconstructions of ”Sofa and cushions”, ”Plants and couch” and ”kitchen and bike” scenes developed using KinectFusion [6, 10] of depth maps generated form StereoSDRNet with refinement. We visualize the camera trajectory, from which the stereo images were taken, via a black curve. Note that for clarity we visualize every 30th frame used by the fusion system.

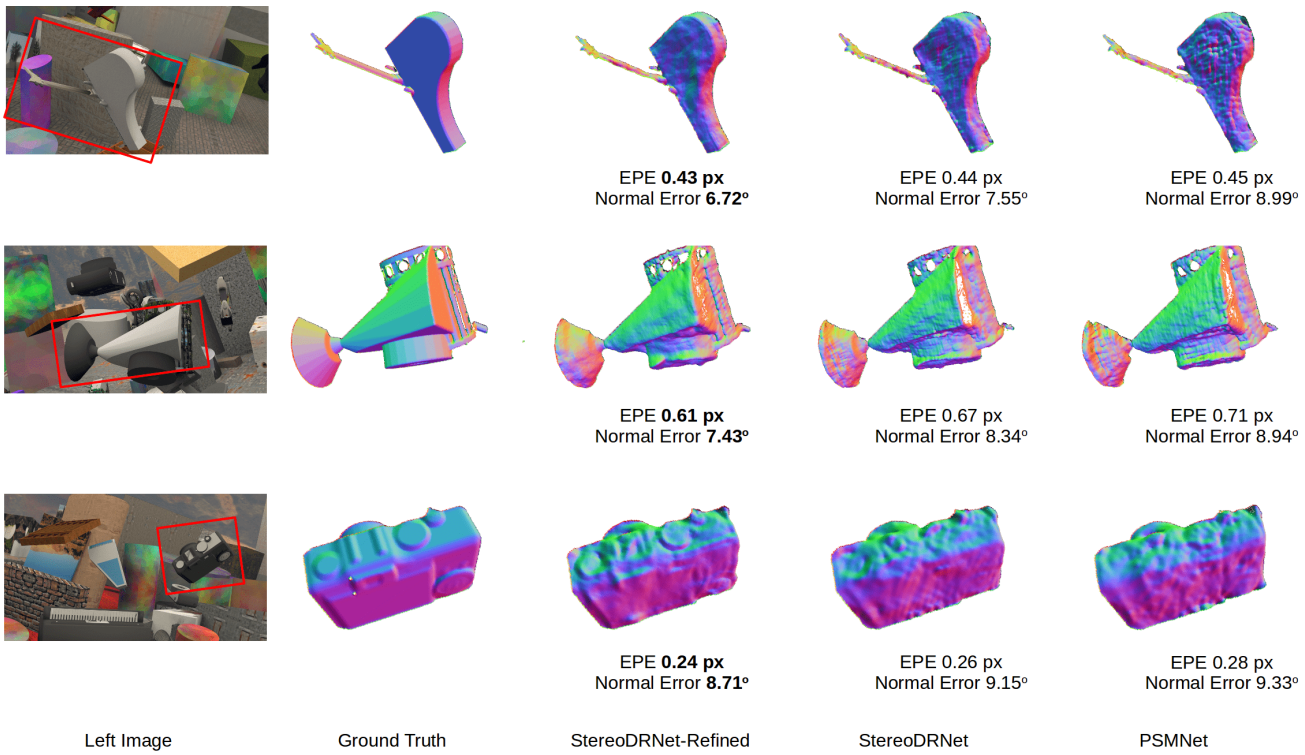

Figure 3: This figure demonstrates the surface normal visualizations of some objects (labeled with red boxes) reconstructed using **single** disparity map from SceneFlow dataset. We report EPE in disparity space and surface normal error in degrees. Notice, our refinement network improves the overall structure of the objects and makes them geometrically consistent.

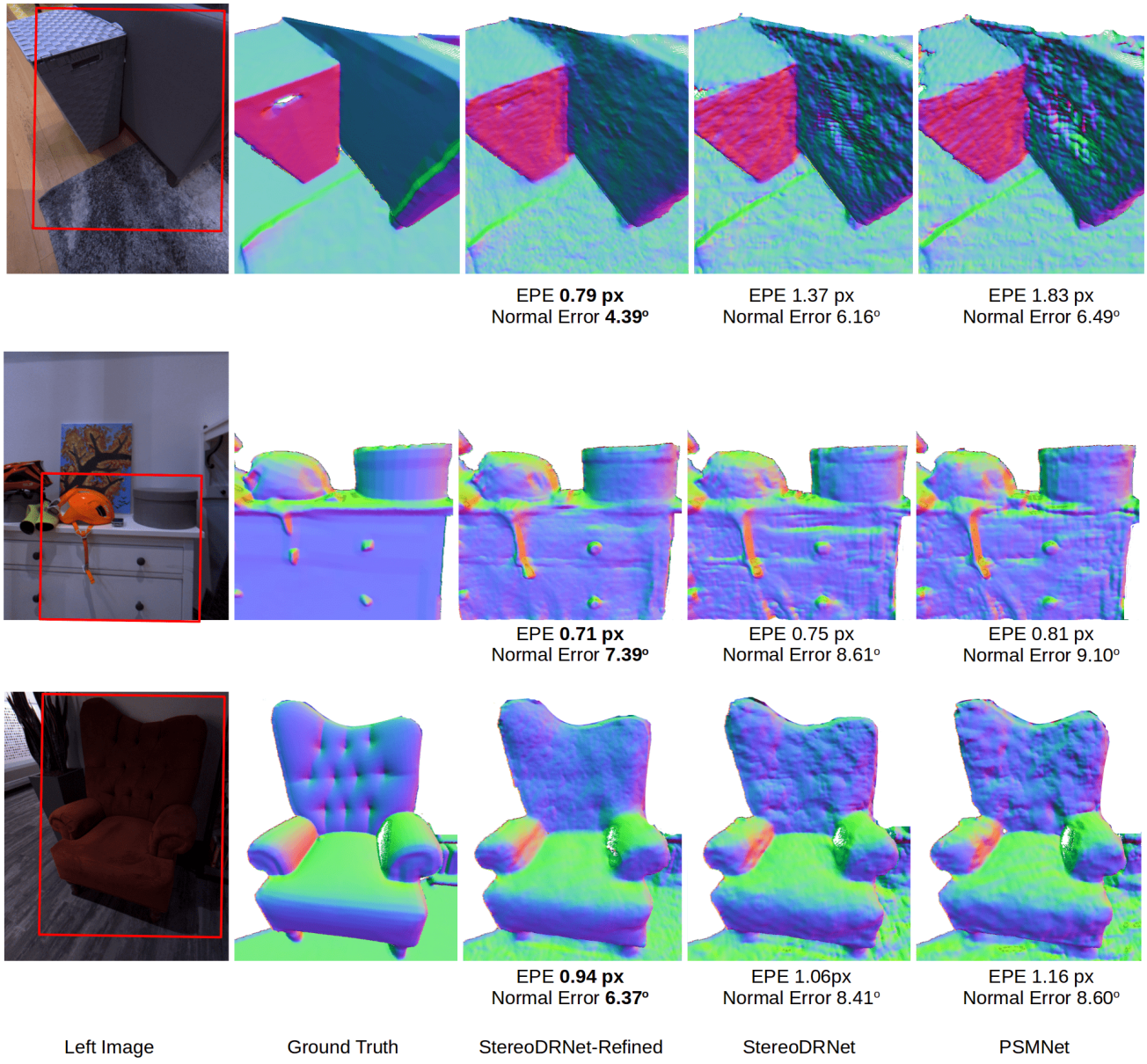

Figure 4: This figure shows the surface normal visualizations of some objects (labeled with red boxes) reconstructed using a **single** disparity map from our real dataset. We report EPE in disparity space and surface normal error in degrees. Notice that our refinement network improves the overall structure of the objects and makes them geometrically consistent.

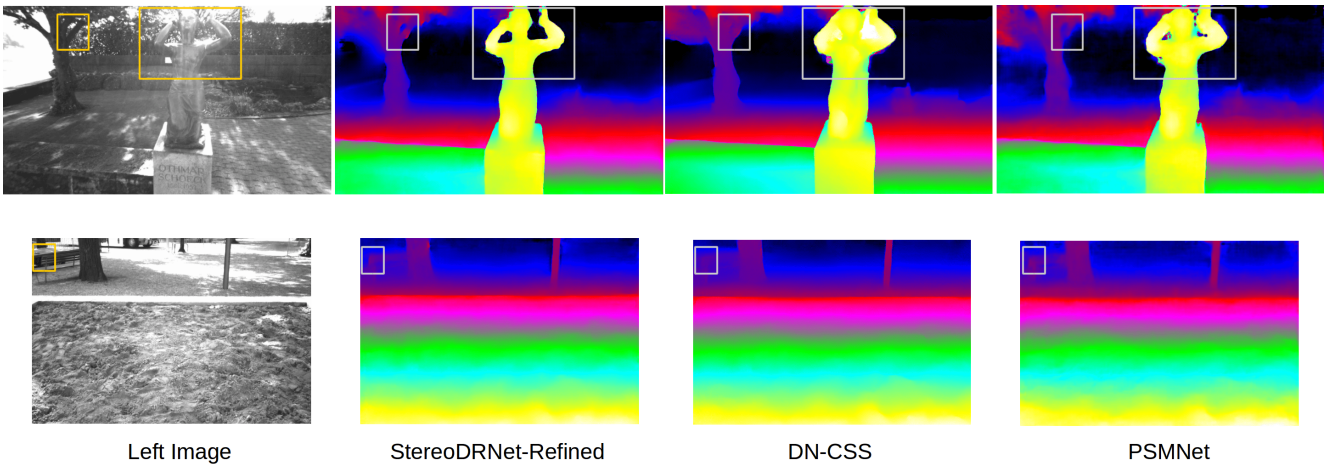

Figure 5: This figure shows the disparity estimation results of our refined network, PSMNet [1] and DN-CSS [2] on the lakeside and sandbox scenes from the ETH3D [8] two view stereo dataset.

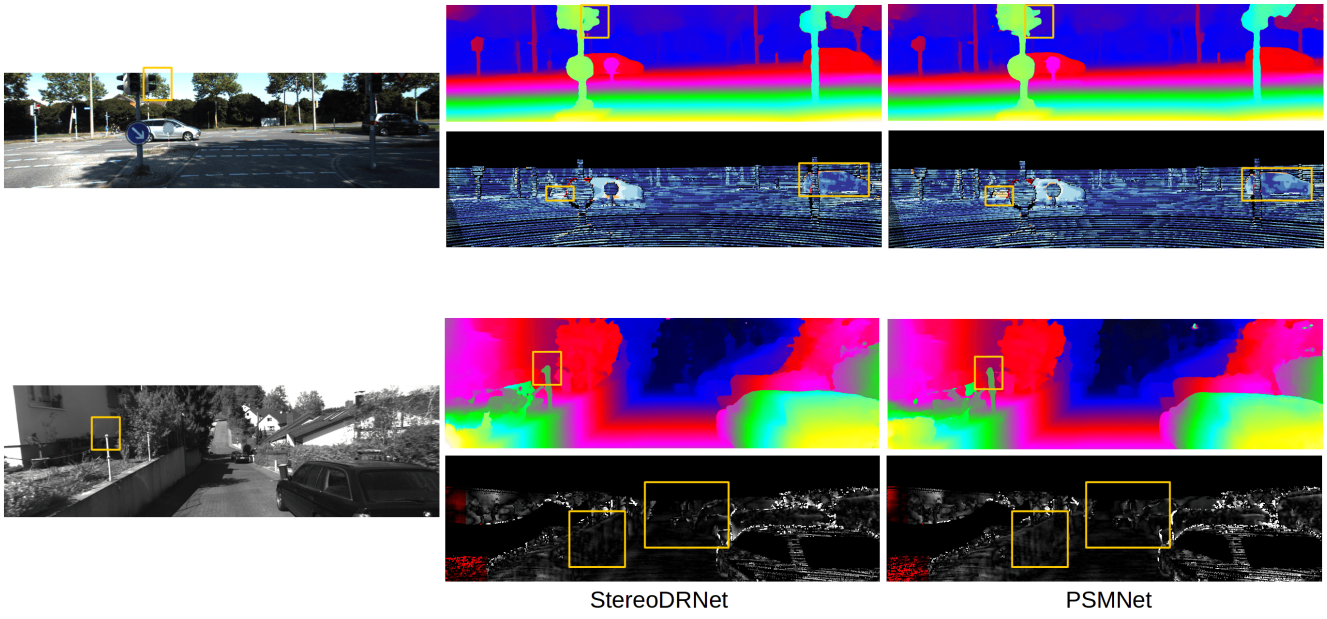

Figure 6: This figure shows the disparity estimation results of our StereoDRNet and PSMNet [1] on the KITTI 2015 and the KITTI 2012 dataset.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 5

[2] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 5

[3] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. *arXiv preprint arXiv:1807.08865*, 2018. 1

[4] Zhengfa Liang, Yiliu Feng, YGHLW Chen, and LQLZJ Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2811–2820, 2018. 1

[5] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 1

[6] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011. 1, 2, 3

[7] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCV Workshops*, volume 7, 2017. 1

[8] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[9] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 1

[10] Thomas Whelan, Michael Goesele, Steven J Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, and Richard Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM Transactions on Graphics (TOG)*, 37(4):102, 2018. 1, 3

[11] Chen-Wei Xie, Hong-Yu Zhou, and Jianxin Wu. Vortex pooling: Improving context representation in semantic segmentation. *arXiv preprint arXiv:1804.06242*, 2018. 1

[12] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017. 1

[13] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: end-to-end self-supervised learning for active stereo systems. *arXiv preprint arXiv:1807.06009*, 2018. 1