

StereoDRNet: Dilated Residual Stereo Net

Rohan Chabra^{1†} Julian Straub² Chris Sweeney² Richard Newcombe² Henry Fuchs¹

¹University of North Carolina at Chapel Hill ²Facebook Reality Labs

¹{rohanc, fuchs}@cs.unc.edu ²julian.straub@oculus.com, {sweenychris, richard.newcombe}@fb.com

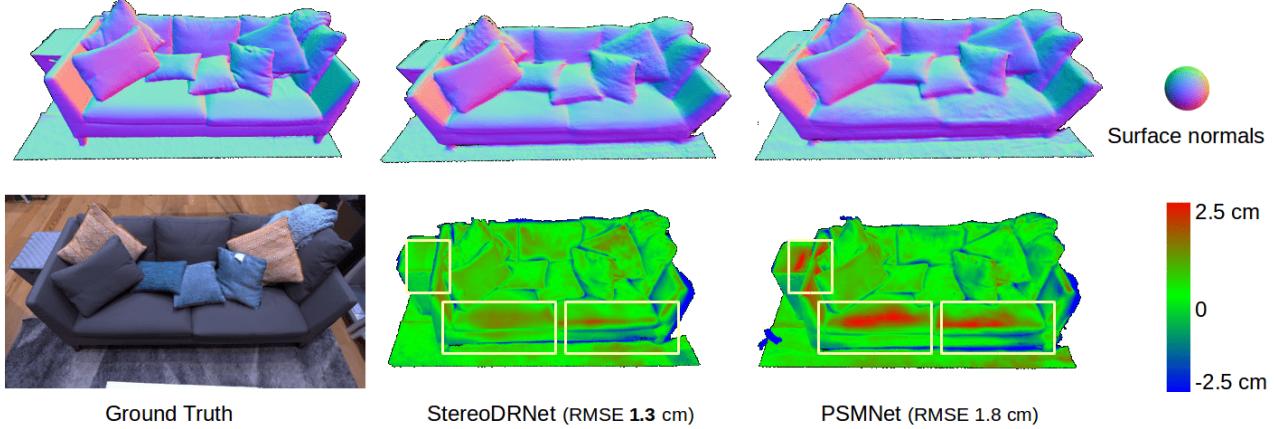


Figure 1: StereoDRNet enables estimation of high quality depth maps that opens the door to high quality reconstruction by passive stereo video. In this figure we compare the output from dense reconstruction [15] built form depth maps generated by StereoDRNet, PSMNet [2] and a structured light system [25] (termed Ground Truth). We report and visualize point-to-plane distance RMS error on the reconstructed meshes with respect to the ground truth demonstrating the improvement in reconstruction over the state-of-the-art. Note that StereoDRNet runs atleast 1.6 times faster than PSMNet.

Abstract

We propose a system that uses a convolution neural network (CNN) to estimate depth from a stereo pair followed by volumetric fusion of the predicted depth maps to produce a 3D reconstruction of a scene. Our proposed depth refinement architecture, predicts view-consistent disparity and occlusion maps that helps the fusion system to produce geometrically consistent reconstructions. We utilize 3D dilated convolutions in our proposed cost filtering network that yields better filtering while almost halving the computational cost in comparison to state of the art cost filtering architectures. For feature extraction we use the Vortex Pooling architecture [26]. The proposed method achieves state of the art results in KITTI 2012, KITTI 2015 and ETH 3D stereo benchmarks. Finally, we demonstrate that our system is able to produce high fidelity 3D scene reconstructions that outperforms the state of the art stereo system.

1. Introduction

Depth from stereo vision has been heavily studied in computer vision field for the last few decades. Depth es-

timation has various applications in autonomous driving, dense reconstruction and 3D objects and human tracking. Virtual Reality and Augmented Reality systems require depth estimations to build dense spatial maps of the environment for interaction and scene understanding. For proper rendering and interaction between virtual and real objects in an augmented 3D world, the depth is expected to be both dense and correct around object boundaries. Depth sensors such as structured light and time of flight sensors are often used to build such spatial maps of indoor environments. These sensors often use illumination sources which require power and space that exceeds the expected budget of an envisioned AR system. Since these sensors use infrared vision, they do not work well in bright sun light environment or in presence of other infrared sources.

On the other hand, the depth from stereo vision systems have a strong advantage of working in both indoors and in sunlight environments. Since these systems use passive image data, they do not interfere with each other or with the environment materials. Moreover, the resolution of passive stereo systems is typically greater than the sparse patterns used in structured light depth sensors, so these methods have capabilities to produce depth with accurate ob-

[†]Work performed during internship at Facebook Reality Labs.

ject boundaries and corners. Due to recent advancements in camera and mobile technology the image sensors have dramatically reduced in size and have significantly improved in resolution and image quality. All these qualities makes passive stereo system a better fit for being a depth estimator for a AR or VR system. However, stereo systems have their own disadvantages, such as ambiguous predictions in texture-less or repeating/confusing textured surfaces. In order to deal with these homogeneous regions traditional methods make use of handcrafted functions and optimize the parameters globally on the entire image. Recent methods use machine learning to derive the functions and it's parameters from the data that is used in training. As these functions tend to be highly non-linear, they tend to yield reasonable approximations even on the homogeneous and reflective surfaces.

Our key contributions are as follows:

- **Novel Disparity Refinement Network:** The main motivation of our work is to predict geometrically consistent disparity maps for stereo input that can be directly used by TSDF-based fusion system like KinectFusion [15] for simultaneous tracking and mapping. Surface normals are an important factor in fusion weight computation in KinectFusion-like systems, and we observed that state of the art stereo systems such as PSMNet produces disparity maps that are not geometrically consistent which negatively affect TSDF fusion. To address this issue, we propose a novel refinement network which takes geometric error E_g , photometric error E_p and unrefined disparity as input and produces refined disparity (via residual learning) and the occlusion map.

- **3D Dilated Convolutions in Cost Filtering:** State of the art stereo systems such as PSMNet[2] and GC-Net[7] that use 3D cost filtering approach use most of the computational resources in the filtering module of their system. We observe that using 3D dilated convolutions in all three dimensions i.e (width, height, and disparity channels) in a structure shown in Fig. 4 gave us better results with less compute (refer to Table.1).

- **Other Contributions:** We observe that Vortex Pooling compared to spatial pyramid pooling (used in PSMNet) provides better results (refer to ablation study 2). We found the exclusion masks used to filter non-confident regions of ground truth for fine-tuning our model as discussed in Sec 4.4 to be very useful in obtaining sharp edges and fine details in disparity predictions. We achieve 1.3 - 2.1 cm RMSE on 3D reconstructions of three scenes that we prepared using structured light system proposed in [25].

2. Related Work

Depth from stereo has been widely explored in the literature, we refer interested readers to surveys and meth-

ods described in [20]. Broadly speaking stereo matching can be categorized into computation of cost metrics, cost aggregation, global or semi-global optimization [4] and refinement or filtering processes. Traditionally global cost filtering approaches used discrete labeling methods such as Graph Cuts [11] or used belief propagation techniques described in [10] and [1]. Total Variation denoising [19] has been used in cost filtering by methods described in [29], [16] and [14].

The state of the art in disparity estimation techniques use CNNs. MC-CNN [30] introduced a Siamese network to compare two image patches. The scores on matching was used along with the semi-global matching process [4] to predict consistent disparity estimation. DispNet [13] demonstrates an end-to-end disparity estimation neural network with a correlation layer (dot product of features) for stereo volume construction. Liang et al. [12] improved DispNet by introducing novel iterative filtering process. GC-Net [7] introduces a method to filter 4D cost using a 3D cost filtering approach and the soft argmax process to regress depth. PSMNet [2] improved GC-Net by enriching features with better global context using pyramid spatial pooling process. They also show effective use of stacked residual networks in cost filtering process.

Xie et al. [26] introduce vortex pooling which is an improvement of the atrous spatial pooling approach used in Deep lab [3]. Atrous pooling uses convolutions with various dilation steps to increase receptive fields of a CNN filter. The vortex pooling technique uses average pooling in grids of varying dimensions before dilated convolutions to utilize information from the pixels which were not used in bigger dilation steps. The size of average pool grids grows with the increase in dilation size. We use the feature extraction described in Vortex pooling and improve the cost filtering approach described by PSMNet.

Our proposed refinement network takes geometric error E_g , photometric error E_p and unrefined disparity as input and produces refined disparity (via residual learning) and the occlusion map. Refinement procedures proposed in CRL [17], iResNet [12], StereoNet [8] and FlowNet2 [5] only use photometric error (either in image or feature domain) as part of the input in the refinement networks. To the best of our knowledge we are the first to explore the importance of geometric error and occlusion training for disparity refinement.

3. Algorithm

In this section we describe our architecture that predicts disparity for the input stereo pair. Instead of using a generic encoder-decoder CNN we break our algorithm into feature extraction, cost volume filtering and refinement procedures.

3.1. Feature Extraction

The feature extraction starts with a small shared weight Siamese network which takes input as images and encodes the input to a set of features. As these features will be used for stereo matching we want them to have both local and global contextual information. To encode local spatial information in our feature maps we start by downsampling the input by use of convolutions with stride of 2. Instead of having a large 5×5 convolution we use three 3×3 filters where first convolution has stride of 2. We bring the resolution to a fourth by having two of such blocks. In order to encode more contextual information we choose Vortex Pooling [26] on the learned local feature maps Fig. 3. Each of our convolutions are followed by batch normalization and ReLU activation except on the last 3×3 convolution on the spatial pooling output. In order to keep the feature information compact we keep the feature dimension size as 32 throughout the feature extraction process.

3.2. Cost Volume Filtering

We use the features extracted in the previous step to produce a stereo cost volume. While several approaches in the literature ([7],[13]) use concatenation or dot products of the stereo features to obtain the cost volume, we found simple arithmetic difference to be just as effective.

While the simple argmin on the cost should in principle lead to the correct local minimum solution, it has been shown several times in literature [16], [4],[20] that it is common for the solution to have several local minima. Surfaces with homogeneous or repeating texture are particularly prone to this problem. By posing the cost filtering as a deep learning process with multiple convolutions and non-linear activations we attempt to resolve these ambiguities and find the correct local minimum.

We start by processing our cost volume with a $3 \times 3 \times 3$ convolution along the width, height and depth dimensions. We then reduce the resolution of the cost by a convolution with stride of 2 followed by convolutions with dilation 1, 2, 4 in parallel. A convolution on the concatenation of the dilated convolution filters is used to combine the information fetched from varying receptive fields.

Residual learning has been shown to be very effective in disparity refinement process so we propose a cascade of such blocks to iteratively improve the quality of our disparity prediction. We depict the entire cost filtering process as Dilated Residual Cost Filtering in Fig. 4. In this figure notice how our network is designed to produce $k = 3$ disparity maps labeled as d^k .

Our network architecture that supports refinement predicts disparities for both left and right view as separate channels in disparity predictions d^k . Note that we construct the cost for both left and right views and concatenate them before filtering; this ensures that the cost filtering method is

provided with cost information for both views. Please refer to Table. 8 for exact architecture details.

3.3. Disparity Regression

In order to have a differentiable argmax we use soft argmax as proposed by GC-Net [7]. For each pixel i the regressed disparity estimation d_i is defined as a weighted softmax function:

$$d_i = \sum_{d=1}^N d \frac{e^{-C_i(d)}}{\sum_{d'=1}^N e^{-C_i(d')}} , \quad (1)$$

where C_i is the cost at pixel i and N is the maximum disparity. The loss L^k for each of the proposed disparity maps d^k (as shown in Fig. 4) in our dilated residual cost filtering architecture, relies on the Huber loss ρ and is defined as:

$$L^k = \sum_i^M \rho(d_i^k, \hat{d}_i) , \quad (2)$$

where d_i^k and \hat{d}_i are the estimated and ground truth disparity at pixel i , respectively and M is the total number of pixels. The total data loss L_d is defined as:

$$L_d = \sum_{k=1}^3 w^k L^k , \quad (3)$$

where w^k is the weight for each disparity map d^k .

3.4. Disparity Refinement

In order to make the disparity estimation robust to occlusions and view consistency we further optimize the estimate. For brevity we label the third disparity prediction d^3 ($k = 3$) described in Sec. 3.2 for left view as D_l and for right view as D_r . In our refinement network we warp the right image I_r to left view via the warp W and evaluate the image reconstruction error map E_p for the left image I_l as:

$$E_p = |I_l - W(I_r, D_r)| . \quad (4)$$

By warping D_r to the left view and using the left disparity D_l we can evaluate the geometric consistency error map E_g as:

$$E_g = |D_l - W(D_r, D_l)| . \quad (5)$$

While we could just reduce these error terms directly into a loss function, we observed significant improvement by using photo-metric and geometric consistency error maps as input to the refinement network as these error terms are only meaningful for non occluding pixels (only pixels for which the consistency errors can be reduced).

Our refinement network takes as input left image I_l , left disparity map D_l , image reconstruction error map E_p and

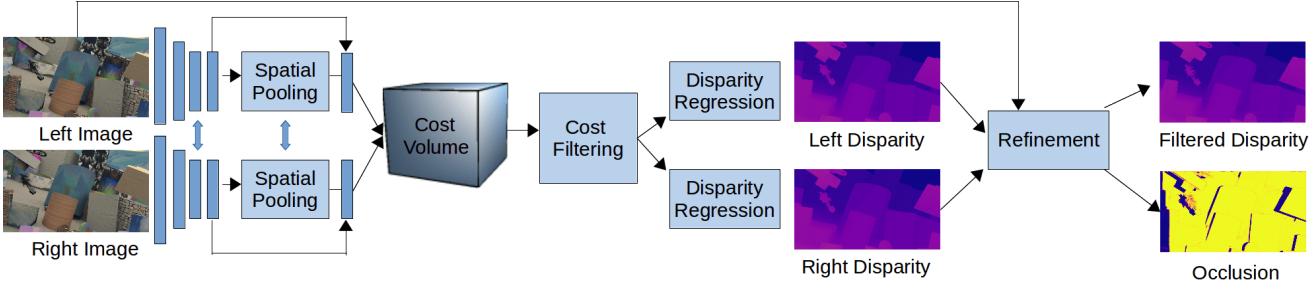


Figure 2: StereoDRNet network architecture pipeline.

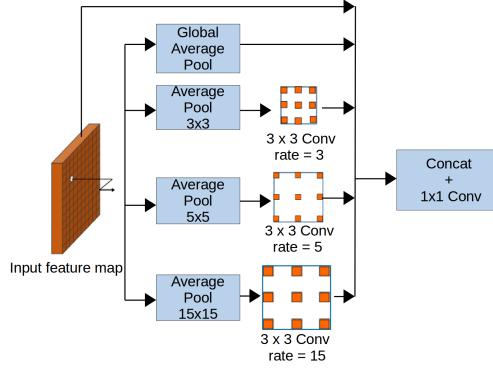


Figure 3: StereoDRNet Vortex Pooling architecture derived from [26].

geometric error map E_g . We first filter left image and reconstruction error and left disparity and geometric error map E_g independently by using one layer of convolution followed by batch normalization. Both these results are then concatenated and followed by atrous convolution [18] to sample from a larger context without increasing the network size. We used dilations with rate 1, 2, 4, 8, 1, and 1 respectively. Finally a single 3×3 convolution without ReLU or batch normalization is used to output an occlusion map O and a disparity residual map R . Our final refined disparity map is labeled as D_{ref} . We demonstrate our refinement network in Fig. 5 and provide exact architecture details in Table. 7.

We compute the cross entropy loss on the occlusion map O as L_o

$$L_o = H(O, \hat{O}), \quad (6)$$

where \hat{O} is the ground truth occlusion map.

The refinement loss L_r is defined as

$$L_r = \sum_i^M \rho(d_i^r, \hat{d}_i), \quad (7)$$

where d_i^r is the value for a pixel i in our refined disparity map D_{ref} and M is the total number of pixels.

Our total loss function L is defined as

$$L = L_d + \lambda_1 L_r + \lambda_2 L_o, \quad (8)$$

where λ_1 and λ_2 are scalar weights.

3.5. Training

We implemented our neural network code in PyTorch. We tried to keep the training of our neural network similar to one described in PSMNet [2] for ease of comparison. We used Adam optimizer [9] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and normalized the image data before passing it to the network. In order to optimize the training procedure we cropped the images to 512x256 resolution. For training we used a mini-batch size of 8 on 2 Nvidia Titan-Xp GPUs. We used $w^1 = 0.2$, $w^2 = 0.4$, $w^3 = 0.6$, $\lambda_1 = 1.2$ and $\lambda_2 = 0.3$ weights in our proposed loss functions Eq. 3 and Eq. 8.

4. Experiments

We tested our architecture on rectified stereo datasets such as SceneFlow, KITTI 2012, KITTI 2015 and ETH3D. We also demonstrate the utility of our system in building 3D reconstruction of indoor scenes. See the supplementary section for additional visual comparisons.

4.1. SceneFlow Dataset

SceneFlow [13] is a synthetic dataset with over 30,000 stereo pairs for training and around 4000 stereo pairs for evaluation. We use both left and right ground truth disparities for training our network. We compute the ground truth occlusion map by defining as occluded any pixel with disparities inconsistency larger than 1 px. This dataset is challenging due to presence of occlusions, thin structures and large disparities.

In Fig. 6 we visually compare our results with PSMNet [2]. Our system infers better structural details in the disparity image and also produces consistent depth maps with significantly less errors in homogeneous regions. We further visualize the effect of our refinement network in Fig. 13.

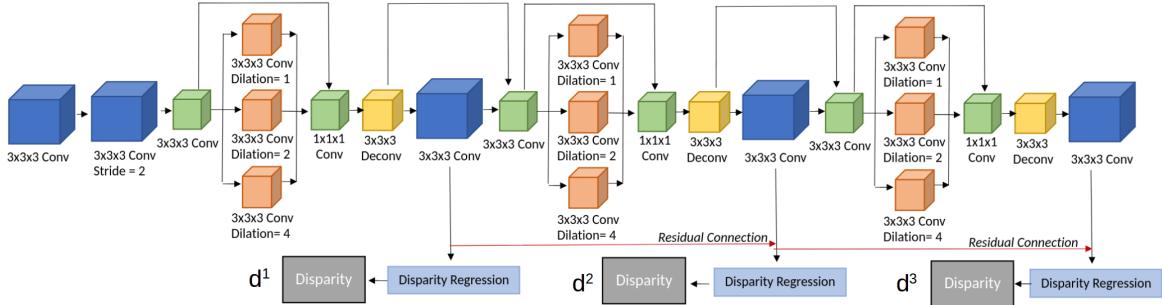


Figure 4: Proposed dilated cost filtering approach with residual connections.

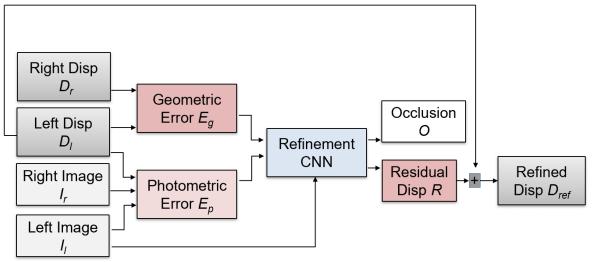


Figure 5: StereoDRNet refinement architecture.

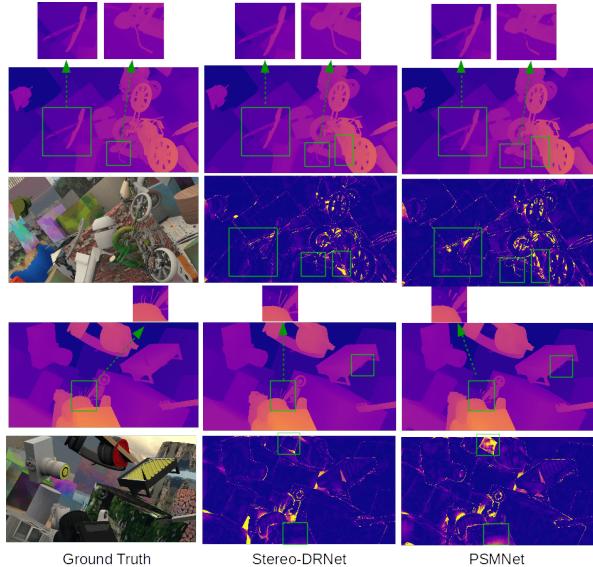


Figure 6: Disparity prediction comparison between our network (Stereo-DRNet) and PSMNet [2] on the SceneFlow dataset. The top row shows disparity and the bottom row shows the EPE map. Note how our network is able to recover thin and small structures and at the same times shows lower error in homogeneous regions.

Table 1 shows a quantitative analysis of our architecture with and without refinement network. Stereo-DRNet achieves significantly lower end point error while reducing computation time. Our proposed cost filtering approach achieves better accuracy with significantly less compute, demonstrating the effectiveness of the proposed dilated residual cost filtering approach.

Ablation study: In Table 2 we show a complete EPE breakdown for different parts of our network on the SceneFlow dataset. Both vortex pooling and refinement procedure add marginal performance gains. Co-training occlusion map with residual disparity drastically improves the mean end point disparity error of the final disparity from 0.93 px to 0.86 px. Passing only the photometric error into the refinement network actually degrades the performance.

Method	EPE	Total FLOPS	3D-Conv FLOPS	FPS
CRL[17]	1.32	-	-	2.1
GC-Net[7]	2.51	8789 GMac	8749 GMac	1.1
PSMNet[2]	1.09	2594 GMac	2362 GMac	2.3
Ours	0.98	1410 GMac	1119 GMac	4.3
Ours-Ref	0.86	1711 GMacs	1356 GMacs	3.6

Table 1: Quantitative comparison of the proposed Stereo-DRNet with the state of the art methods on the SceneFlow dataset. EPE represent the mean end point error in disparity. FPS and FLOPS (needed by the convolution layers) are measured on full 960×540 resolution stereo pairs. Notice even our unrefined disparity architecture outperforms the state of the art method PSMNet [2] while requiring significantly less computation.

4.2. KITTI Datasets

We evaluated our method on both KITTI 2015 and KITTI 2012 datasets. These data sets contain stereo pairs with semi-dense depth images acquired using a LIDAR sensor that can be used for training. The KITTI 2012 dataset contains 194 training and 193 test stereo image pairs from static outdoor scenes. The KITTI 2015 dataset contains 200 training and 200 test stereo image pairs from both static and

Network Architecture				SceneFlow	KITTI-2015		
Pooling	Cost Filtering			Refinement		EPE	Val Error(%)
	d^1	d^2	d^3	E_p	E_g		
Pyramid	✓					1.17	2.28
Vortex	✓					1.13	2.14
Vortex	✓	✓				0.99	1.88
Vortex	✓	✓	✓			0.98	1.74
Pyramid	✓	✓	✓			1.00	1.81
Vortex	✓	✓	✓	✓		1.03	-
Vortex	✓	✓	✓		✓	0.95	-
Vortex	✓	✓	✓	✓	✓	0.93	-
Vortex	✓	✓	✓	✓	✓	0.86	-
Pyramid	✓	✓	✓	✓	✓	0.96	-

Table 2: Ablation study of network architecture settings on SceneFlow and KITTI-2015 evaluation dataset.

Method	2px		3px		Avg Error		Time(s)
	Noc	All	Noc	All	Noc	All	
GC-NET[7]	2.71	3.46	1.77	2.30	0.6	0.7	0.90
EdgeStereo[22]	2.79	3.43	1.73	2.18	0.5	0.6	0.48
PDSNet[24]	3.82	4.65	1.92	2.53	0.9	1.0	0.50
SegStereo[27]	2.66	3.19	1.68	2.03	0.5	0.6	0.60
PSMNet[2]	2.44	3.01	1.49	1.89	0.5	0.6	0.41
Ours	2.29	2.87	1.42	1.83	0.5	0.5	0.23

Table 3: Comparison of disparity estimation from Stereo-DRNet with state of the art published methods on KITTI 2012 dataset.

Method	All(%)			Noc(%)			Time(s)
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
DN-CSS[6]	2.39	5.71	2.94	2.23	4.96	2.68	0.07
GC-NET[7]	2.21	6.16	2.87	2.02	5.58	2.61	0.90
CRL[17]	2.48	3.59	2.67	2.32	3.12	2.45	0.47
EdgeStereo[22]	2.27	4.18	2.59	2.12	3.85	2.40	0.27
PDSNet[24]	2.29	4.05	2.58	2.09	3.69	2.36	0.50
PSMNet[2]	1.86	4.62	2.32	1.71	4.31	2.14	0.41
SegStereo[27]	1.88	4.07	2.25	1.76	3.70	2.08	0.60
Ours	1.72	4.95	2.26	1.57	4.58	2.06	0.23

Table 4: Comparison of disparity estimation from Stereo-DRNet with state of the art published methods on KITTI 2015 dataset.

dynamic outdoor scenes.

Training and ablation study: Since KITTI data sets contain only limited amount of training data, we fine tuned our model on the SceneFlow dataset. In our training we used 80% stereo pairs for training and 20% stereo pairs for evaluation. We demonstrate the ablation study of our proposed method on KITTI 2015 dataset Table 2. Note how our proposed dilated residual architecture and the use of Vortex pooling for feature extraction consistently improve the re-

Method	All				Noc			
	1px	2px	4px	RMSE	1px	2px	4px	RMSE
PSMNet[2]	5.41	1.31	0.54	0.75	5.02	1.09	0.41	0.66
iResNet[12]	4.04	1.20	0.34	0.59	3.68	1.00	0.25	0.51
DN-CSS[6]	3.00	0.96	0.34	0.56	2.69	0.77	0.26	0.48
Ours	4.84	0.96	0.30	0.55	4.46	0.83	0.24	0.50

Table 5: Comparison of disparity estimation from Stereo-DRNet with state of the art published methods on ETH 3D dataset.

sults. We did not achieve significant gains by doing refinement on KITTI datasets as these datasets only contain labeled depth for sparse pixels. Our refinement procedure improves disparity predictions using view consistency checks and sparsity in ground truth data affected the training procedure. We demonstrate that data sets with denser training data enabled the training and fine-tuning of our refinement model.

Results: We evaluated our Dilated residual network without filtering on both these datasets and achieved state of the art results on KITTI 2012 Table 3 and comparable results with best published method on KITTI 2015 Table 4. On KITTI 2015 dataset the three columns D1-bg, D1-fg and D1-all mean that the pixels in the background, foreground, and all areas, respectively, were considered in the estimation of errors. We perform consistently well in D1-bg meaning background areas, we achieve comparable results with state of art method in all pixels and better results in non-occluded regions. On KITTI 2012 dataset "Noc" means non occluded regions and "All" mean all regions. Notice, that we perform comparable against SegStereo [27] on KITTI 2015 but way better in KITTI 2012 dataset.

4.3. ETH3D Dataset

We again used our pre-trained network trained on Scene-flow dataset and fine-tuned it on the training set provided in the dataset. ETH dataset contains challenging scenes of both outside and indoor environment. According to our Table 5 we perform best on almost half of the evaluation metrics, our major competitor in this evaluation was DN-CSS [6]. Although, we observe that this method did not perform well on KITTI 2015 data set Table 4. Notice, as this data set contained dense training disparity maps of both stereo views we were able to train and evaluate our refinement network on this data set.

4.4. Indoor Scene Reconstruction

We use the scanning rig used in recent work [25] for preparing ground truth dataset for supervised learning of depth and added one more RGB camera to the rig to obtain a stereo image pair. We kept the baseline of the stereo pair to be about 10cm. We trained our StereoDRNet network on

SceneFlow as described in section 4.1 and then fine tuned the pre-trained network on 250 stereo pairs collected in the indoor area by our scanning rig. We observed that the network quickly adapted to our stereo rig with a minimal amount of fine-tuning.

For preparing ground truth depth we found rendered depth from complete scene reconstruction to be a better estimate than the live sensor depth which usually suffers from occlusions and depth uncertainties. Truncated signed distance function (TSDF) was used to fuse live depth maps into a scene as described in [15].

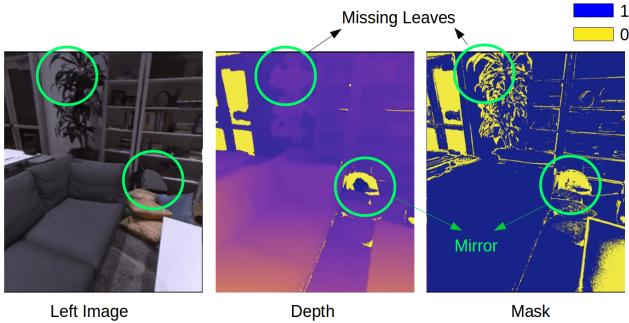


Figure 7: We show a training example with the left image, ground truth depth and the exclusion mask. Note that the glass, mirrors and the sharp corners of the table are excluded from training as indicated by the yellow pixels in the occlusion mask. Note, that this example was not part of our actual training set.

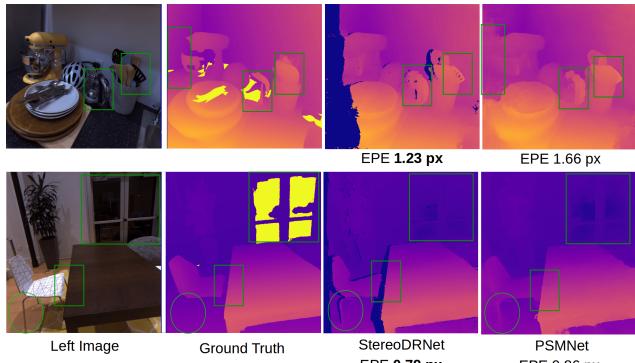


Figure 8: This figure demonstrates that our StereoDRNet network produces better predictions on thin reflective legs of the chair and some portions of the glass. We used occlusion mask predicted by our network to clip occluding regions. Yellow region in the ground truth are the regions that belong to our proposed exclusion mask.

The infrared-structure light depth sensors are known to be unresponsive to dark and highly reflective surfaces. Moreover, the quality of TSDF fusion is limited to the res-

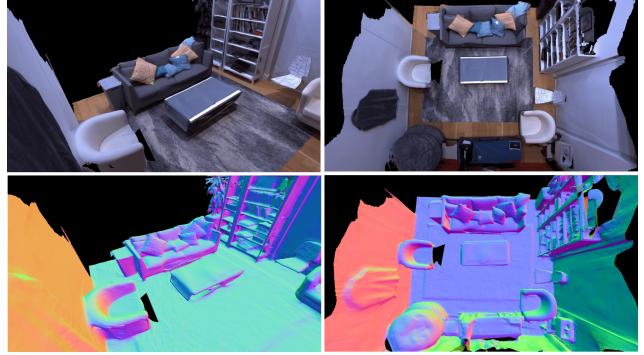


Figure 9: This figure demonstrates 3D reconstruction of a living room in an apartment prepared by TSDF fusion of the predicted depth maps from our system. We visualize two views of the textured mesh and surface normals in top and bottom rows respectively.

olution of the voxel size. Hence we expect the reconstructions to be overly smooth in some areas such as table corners or sharp edges of plant leaves. In order to avoid contaminating our training data with false depth estimation, we use a simple photometric error threshold to mask out the pixels from training where the textured model projection color disagrees with the real images. We show one such example in Fig. 7 where glass, mirrors and the sharp corners of the table are excluded from training. Although, the system from Whelan et al. [25] can obtain ground truth planes of mirrors and glass we avoid depth supervision on them in this work as it is beyond the scope of a stereo matching procedure to obtain depth on reflectors.

We demonstrate visualizations of the depth predictions from the stereo pair in Fig. 8. Notice, our prediction is able to recover sharp corners of the table, thin reflective legs of the chair and several thin structures in kitchen dataset as a result of filtering process used in training. It is interesting to see that we recover the top part of the glass correctly but not the bottom part of the glass which suffers from reflections. The stereo matching model simply treats reflectors as windows in presence of reflections.

Results and evaluations: We demonstrate visualizations of full 3D reconstruction of a living room in an apartment prepared by TSDF fusion of the predicted depth maps from our system in Fig. 9. For evaluation study we prepared three small data sets that we refer as ‘‘Sofa and cushions’’ demonstrated in Fig. 1, ‘‘Plants and couch’’ and ‘‘Kitchen and bike’’ demonstrated in Fig. 10. We report point-to-plane root mean squared error (RMSE) of the reconstructed 3D meshes from fusion of depth maps obtained from PSM-Net [2] and our refined network. We obtain a RMSE of 1.3 cm on the simpler ‘‘Sofa and cushions’’ dataset. Note that our method captured high frequency structural details

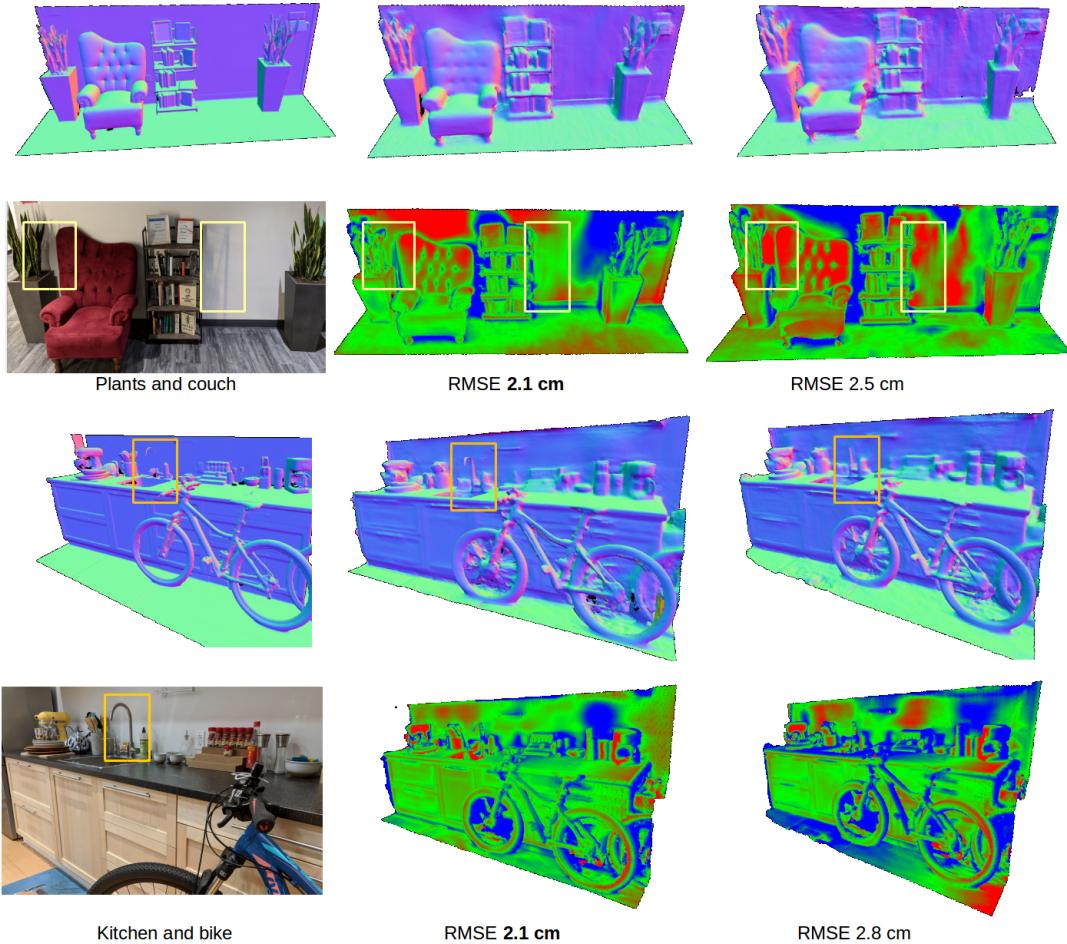


Figure 10: Comparison of 3D reconstruction using fusion of depth maps from our StereoDRNet network (middle), PSMNet [2] (right) and depth maps from the structured light system (left) described in [25] (termed Ground Truth). We report and visualize point-to-plane distance RMS error on the reconstructed meshes with respect to the ground truth mesh. Dark yellow boxes represent the regions where our reconstruction yields details that the structured light sensor or PSMNet were not able to capture. Light yellow boxes represent regions where StereoDRNet outperforms PSMNet.

on the cushions which were not captured by PSMNet or the structured light sensor. ‘‘Plants and couch’’ represents a more difficult scene as it contained a directed light source casting shadows. For this dataset StereoDRNet obtained 2.1 cm RMSE whereas PSMNet obtained 2.5 cm RMSE. Notice, that our reconstruction is not only cleaner but produces minimal errors in the shadowed areas (shadows cast by book shelf and left plant). ‘‘Kitchen and bike’’ dataset cluttered and contains reflective objects making it the hardest dataset. While our system still achieved 2.1 cm RMSE, the performance of PSMNet degraded to 2.8 cm RMSE. Notice, that our reconstruction contains the faucet (highlighted by yellow box) in contrast to the structured light sensor and PSMNet reconstructions. For all evaluations we used exactly the same training dataset for fine-tuning our StereoDRNet and PSMNet.

5. Conclusion

Depth estimation from passive stereo images is a challenging task. Systems from related work suffer in regions with homogeneous texture or surfaces with shadows and specular reflections. Our proposed network architecture uses global spatial pooling and dilated residual cost filtering techniques to approximate the underlying geometry even in above mentioned challenging scenarios. Furthermore, our refinement network produces geometrically consistent disparity maps with the help of occlusion and view consistency cues. The use of perfect synthetic data and careful filtering of real training data enabled us to recover thin structures and sharp object boundaries. Finally, we demonstrate that our passive stereo system, when used for building 3D scene reconstructions in challenging indoor scenes, approaches the quality of state-of-the-art structured light systems [25].

Supplementary

A. Overview

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 11: 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 [25].

B. 3D Reconstruction Experiments

For all 3D reconstruction experiments and evaluations we used a set of about 200 stereo views shown in Fig. 11 to fine tune the SceneFlow [13]-pre-trained networks.

We show the textured 3D reconstructions of our indoor scene dataset in Fig. 12. Note that we used KinectFusion [15] 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. 12 are the output of the ICP (iterative closest point) procedure used by the KinectFusion [15] system. We used manual adjustment followed by ICP to align the 3D reconstructions wherever necessary for our evaluations.

C. Network Details

We provide the network architecture of StereoDRNet in Table. 8. We borrowed ideas on extracting robust local image features from PSMNet [2]. As described in the paper, we use Vortex Pooling [26] 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 6: 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. 6 on the SceneFlow [13] dataset. We observed that increasing dilation rates improved the quality of predictions. Dilation rates above 4 did not provide any significant gains.

Index	Layer Description	Output
1	Warp(I_R, d_L^3) - I_L	H x W x 3
2	concat 1, I_L	H x W x 6
3	Warp(d_R^3, d_L^3) - d_L^3	H x W x 1
4	concat 3, 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	d^r : 14(a) + d_L^3	H x W
16	O : sigmoid on 14(b)	H x W

Table 7: Refinement network for StereoDRNet. d^r and O represent refined disparity and occlusion probability respectively.

The proposed refinement network described in Table. 7 is inspired by the refinement procedures proposed in CRL [17], iResNet [12], StereoNet [8], and ActiveStereoNet [31]. We adopted the basic architecture for refinement as described in StereoNet [8] with dilated residual blocks [28] to increase the receptive field of filtering without compromising resolution. This technique was also adopted in recent work on optical flow prediction Pwc-

net [23]. We experienced additional gains when using the photometric error E_p and geometric error maps E_g as inputs and co-training of occlusion maps. Such enhancements in the refinement procedure has never been proposed to the best of our knowledge.

D. 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 [15]. We use a surface normal error metric to measure consistency in the disparity predictions (first order derivative). Figures 13 and 14 visualize how our refinement procedure improves the overall structure of objects. In some cases such as in the first comparison in Fig. 13 we observe little improvement in disparity prediction but large improvement in surface normals. Figure 14 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 [15] 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 8: 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.

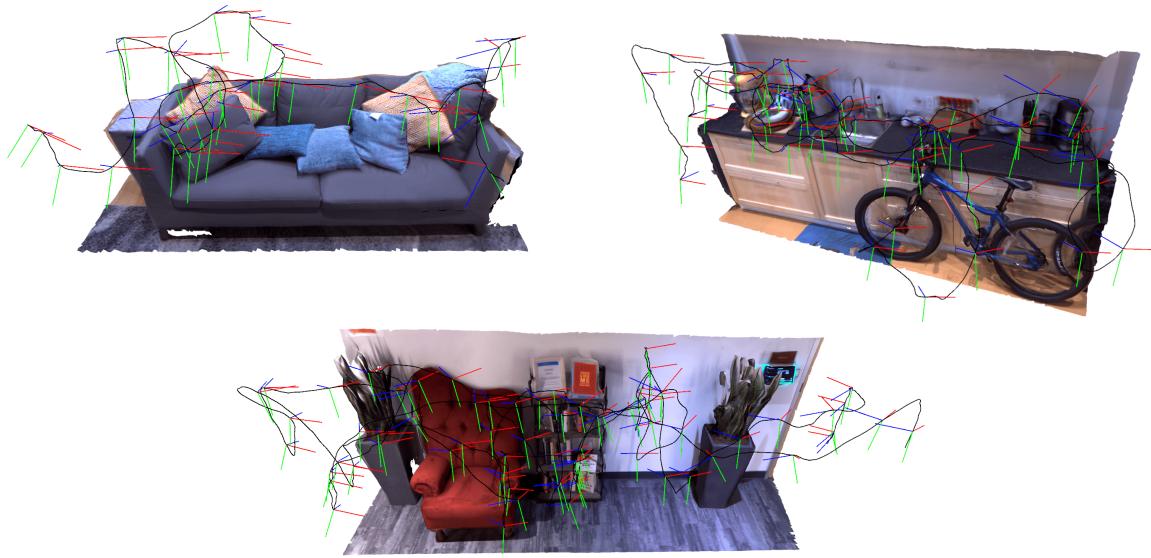


Figure 12: This figure shows the textured 3D reconstructions of "Sofa and cushions", "Plants and couch" and "kitchen and bike" scenes developed using KinectFusion [15, 25] of depth maps generated from StereoSDRNet with refinement. We visualize the camera trajectory, from which the stereo images were taken, via a black curve. Note that for clarity we visualize every 30th frame used by the fusion system.

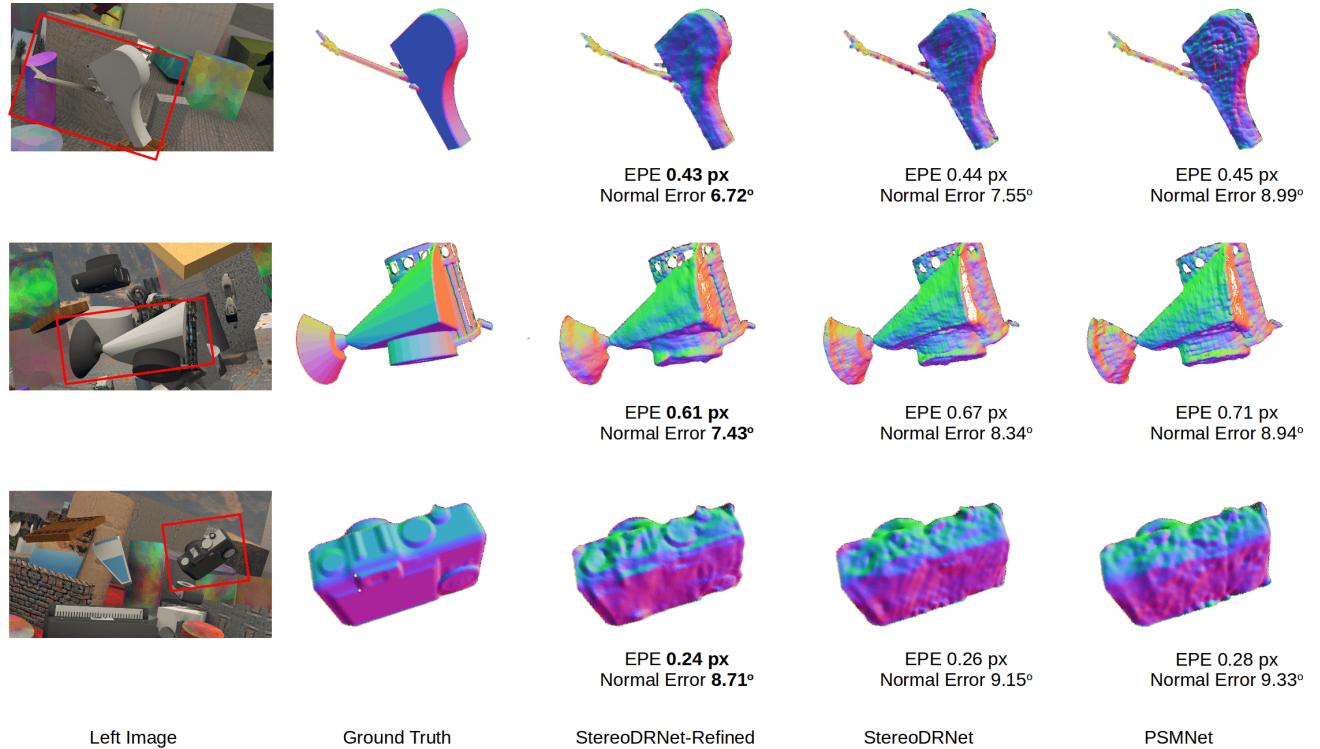


Figure 13: This figure demonstrates the surface normal visualizations of some objects (labeled with red boxes) reconstructed using **single** disparity map from SceneFlow dataset. We report EPE in disparity space and surface normal error in degrees. Notice, our refinement network improves the overall structure of the objects and makes them geometrically consistent.

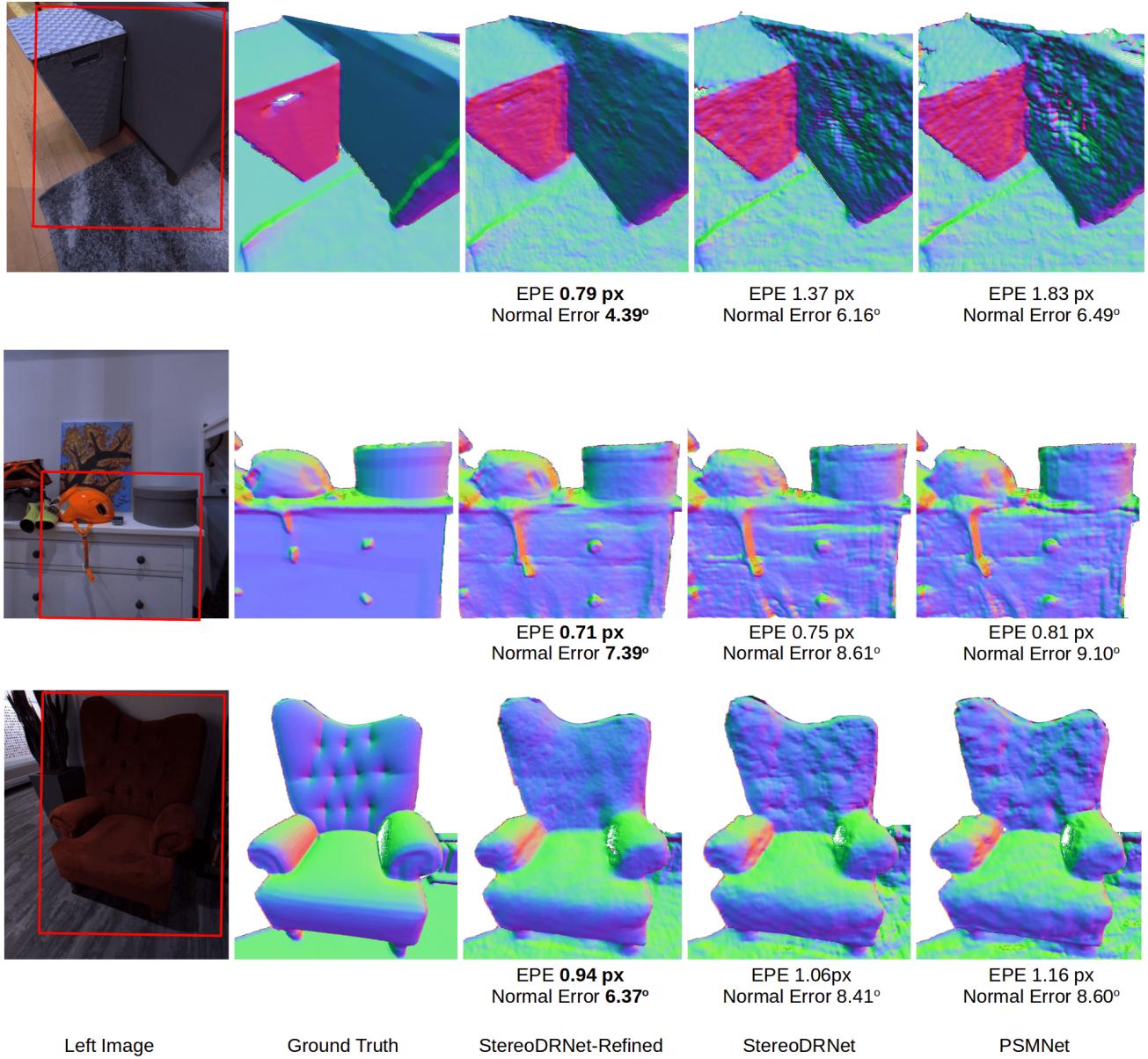


Figure 14: This figure shows the surface normal visualizations of some objects (labeled with red boxes) reconstructed using a **single** disparity map from our real dataset. We report EPE in disparity space and surface normal error in degrees. Notice that our refinement network improves the overall structure of the objects and makes them geometrically consistent.

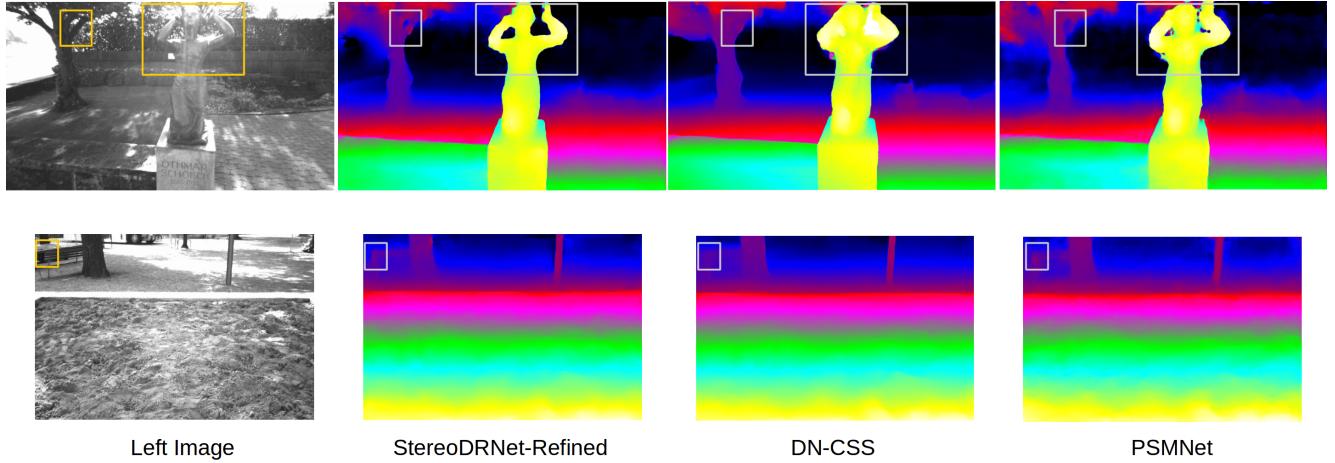


Figure 15: This figure shows the disparity estimation results of our refined network, PSMNet [2] and DN-CSS [6] on the lakeside and sandbox scenes from the ETH3D [21] two view stereo dataset.

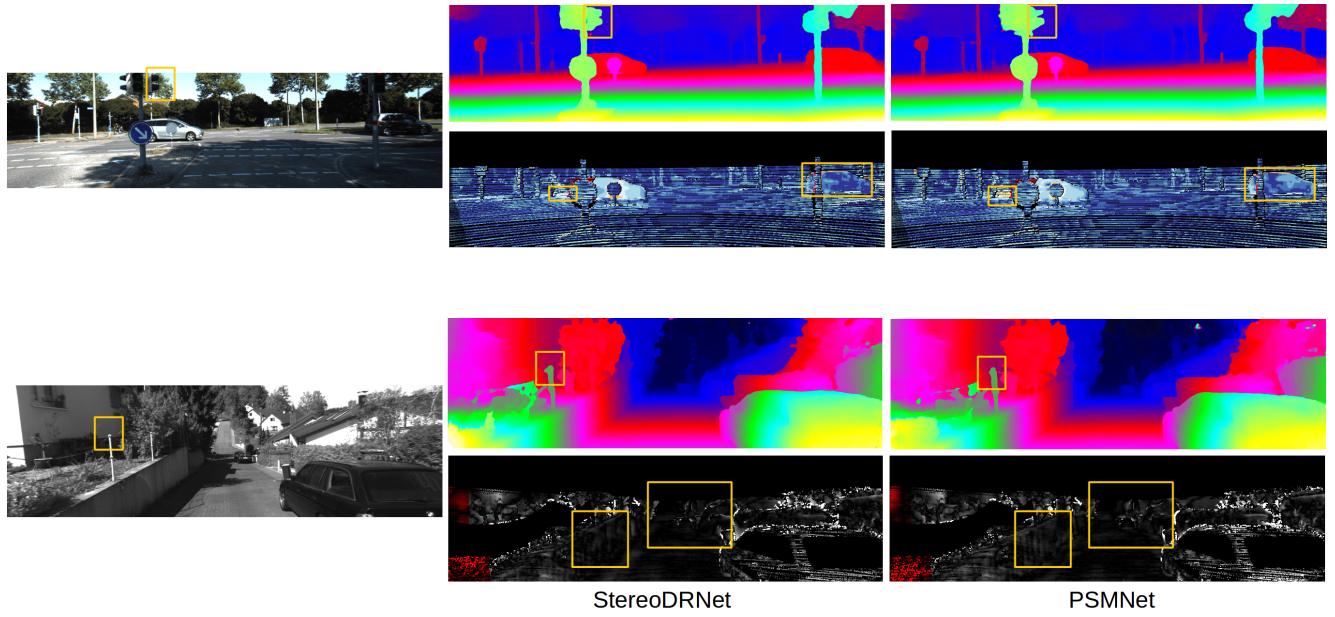


Figure 16: This figure shows the disparity estimation results of our StereoDRNet and PSMNet [2] on the KITTI 2015 and the KITTI 2012 dataset.

References

- [1] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011. 2
- [2] 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, 2, 4, 5, 6, 7, 8, 9, 13
- [3] 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 transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 2
- [4] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008. 2, 3
- [5] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6, 2017. 2
- [6] 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. 6, 13
- [7] 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. *CoRR*, vol. abs/1703.04309, 2017. 2, 3, 5, 6
- [8] 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. 2, 9
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4
- [10] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 15–18. IEEE, 2006. 2
- [11] Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions using graph cuts. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 508–515. IEEE, 2001. 2
- [12] 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. 2, 6, 9
- [13] 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. 2, 3, 4, 9
- [14] Richard Newcombe. *Dense visual SLAM*. PhD thesis, Imperial College London, UK, 2012. 2
- [15] 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, 7, 9, 10, 11
- [16] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011. 2, 3
- [17] 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. 2, 5, 6, 9
- [18] George Papandreou, Iasonas Kokkinos, and Pierre-André Savalle. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 390–399, 2015. 4
- [19] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 2
- [20] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002. 2, 3
- [21] 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. 13
- [22] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. *arXiv preprint arXiv:1803.05196*, 2018. 6
- [23] 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. 10
- [24] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *arXiv preprint arXiv:1806.01677*, 2018. 6
- [25] 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, 2, 6, 7, 8, 9, 11

- [26] Chen-Wei Xie, Hong-Yu Zhou, and Jianxin Wu. Vortex pooling: Improving context representation in semantic segmentation. *arXiv preprint arXiv:1804.06242*, 2018. [1](#), [2](#), [3](#), [4](#), [9](#)
- [27] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. *arXiv preprint arXiv:1807.11699*, 2018. [6](#)
- [28] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017. [9](#)
- [29] Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust tv-l 1 range image integration. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. [2](#)
- [30] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016. [2](#)
- [31] 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. [9](#)