Normal Assisted Stereo Depth Estimation: Supplementary Material

Uday Kusupati[1*] Shuo Cheng[2] Rui Chen[3*] Hao Su[2]

[1]The University of Texas at Austin [2]University of California San Diego

[3]Tsinghua University

uday@cs.utexas.edu, scheng@eng.ucsd.edu, chenr17@mails.tsinghua.edu.cn, haosu@eng.ucsd.edu

1. Implementation details

We use 64 levels of depth/disparity while building the cost volumes. The hyperparameters in the loss function λ_z and λ_n are set to 0.7 and 3 respectively. We train the network without the consistency module first for 30 epochs with ADAM optimizer with a learning rate of 2×10^{-4}. Further, we finetune the consistency module with the end-to-end pipeline for 10 epochs with a learning rate of 1×10^{-4}. The training process takes 5 days and uses 4 NVIDIA GTX 1080Ti GPUs with a batch size of 12. We use a random crop size of (320×240) during training which can be optionally increased in the later epochs by decreasing the batch size.

2. View Selection and Normal Generation

ScanNet [2] provides depth map and camera pose for each image frame. To make it appropriate for stereo evaluation, view selection is a crucial step. Following Yao *et al*. [5], we calculate a score $s(i,j) = \sum_{\mathbf{p}} \mathcal{G}(\theta_{ij}(\mathbf{p}))$ for each image pair according to the sparse points, where $\mathbf{p}$ is a common track in both view i and j, $\theta_{ij}(\mathbf{p}) = (180/\pi)\arccos((\mathbf{c}_i-\mathbf{p})\cdot(\mathbf{c}_j-\mathbf{p}))$ is $\mathbf{p}$'s baseline angle and $\mathbf{c}$ is the camera center. $\mathcal{G}$ is a piece-wise Gaussian function [6] that favors a certain baseline angle θ_0:

$$\mathcal{G}(\theta) = \begin{cases} \exp(-\frac{(\theta-\theta_0)^2}{2\sigma_1^2}), \theta \leq \theta_0 \\ \exp(-\frac{(\theta-\theta_0)^2}{2\sigma_2^2}), \theta > \theta_0 \end{cases}$$

In the experiments, θ_0, σ_1 and σ_2 are set to 5°, 1 and 10 respectively. We generate ground-truth surface normal maps following the procedure of [3].

3. Visualization of NNet slices

We justify the intuition in Section 3.2 in the main paper by visualising the normal estimate contribution from each slice *i.e*. **NNet**(S_i) in Figure 1. The slices in the figure clearly show that only slices with good correspondence probabilities contribute to the output of NNet.

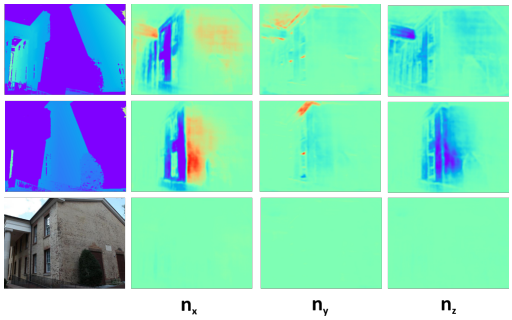

Figure 1. **Normal Estimation contribution from different slices.** The top two rows shows the mask of receptive field and contribution of normal prediction of two slices S_i close to the ground truth depth. The third row shows the sum of the outputs of NNet on all other slices.

4. $\mathcal{L}_t$ and Comparison with $\mathcal{L}_c$

We first analyse the depth propagation method using normals proposed in [4] and reduce it to a form where we can compare it with $\mathcal{L}_c$. In [4], given the depth estimate of pixel i, Z_i is accurate, the depth estimate of neighboring pixel j, Z_j is estimated using,

$$Z_j = \frac{n_x X_j + n_y Y_j + n_z Z_j}{(u_i - c_x)n_x/f_x + (v_i - c_y)n_y/f_y + n_z} \quad (1)$$

where (n_x, n_y, n_z) is the normal map estimate at j. This equation can be rearranged to

$$(Z_i - Z_j) = \frac{Z_i u_i - Z_j u_j}{f_x}\left(\frac{-n_x}{n_z}\right) + \frac{Z_i v_i - Z_j v_j}{f_y}\left(\frac{-n_y}{n_z}\right)$$

$$(Z_i - Z_j) = (X_i - X_j)\left(\frac{-n_x}{n_z}\right) + (Y_i - Y_j)\left(\frac{-n_y}{n_z}\right)$$

$$\Delta Z = \Delta X\left(\frac{-n_x}{n_z}\right) + \Delta Y\left(\frac{-n_y}{n_z}\right)$$

Eq. 6 $\Rightarrow$

$$\Delta Z = \Delta X\left(\frac{\partial Z}{\partial X}\right) + \Delta Y\left(\frac{\partial Z}{\partial Y}\right) \tag{2}$$

From definition of total derivative,

$$dZ = dX\left(\frac{\partial Z}{\partial X}\right) + dY\left(\frac{\partial Z}{\partial Y}\right) \tag{3}$$

In [4], the authors use the assumption that neighboring pixels can be assumed to be lying on the same tangent plane, which is the same as approximating $\left(\frac{dZ}{dX}, \frac{dZ}{dY}\right)$ with $\left(\frac{\Delta Z}{\Delta X}, \frac{\Delta Z}{\Delta Y}\right)$.

We now compare this formulation of depth-normal consistency with ours. Considering neighboring pixels along X-direction, $\frac{\Delta Z}{\Delta X} = \frac{\partial Z}{\partial X}$, and similarly, $\frac{\Delta Z}{\Delta Y} = \frac{\partial Z}{\partial Y}$. This formulation can be put as an objective function minimization,

$$\mathcal{L}_t = \left|\left(\frac{\Delta Z}{\Delta X}, \frac{\Delta Z}{\Delta Y}\right) - \left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}\right)\right|_{\mathbf{H}} \tag{4}$$

Our formulation $\mathcal{L}_c$ is,

$$\mathcal{L}_c = \left|\left(\frac{\Delta Z}{\Delta u}, \frac{\Delta Z}{\Delta v}\right) - \left(\frac{\partial Z}{\partial u}, \frac{\partial Z}{\partial v}\right)\right|_{\mathbf{H}} \tag{5}$$

So fundamentally, while previous depth-normal consistencies generally deal in world coordinate space, we concentrate on pixel coordinate space, because the depth map we estimate is a function $Z(u, v)$ in u, v. By minimizing $\mathcal{L}_c$, we make the assumption of approximating $\left(\frac{\partial Z}{\partial u}, \frac{\partial Z}{\partial v}\right)$ with $\left(\frac{\Delta Z}{\Delta u}, \frac{\Delta Z}{\Delta v}\right)$, in contrast to approximating $\left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}\right)$ with $\left(\frac{\Delta Z}{\Delta X}, \frac{\Delta Z}{\Delta Y}\right)$. The first formulation $\mathcal{L}_t$ enforces depth gradient consistency in world coordinate space with the assumption that the depth gradients are locally linear in world coordinate space. The second formulation $\mathcal{L}_c$ enforces depth gradient consistency in pixel coordinate space with the assumption that the depth gradients are locally linear in pixel coordinate space.

Due to the camera projection geometry, the separation between world coordinates of neighboring pixels in X and Y directions depends on the absolute depth at the pixels. The depth gradient linearity assumption in the world coordinate space, assumes the depth gradient to be locally linear at all depth scales, irrespective of the absolute depth. Where as, in our formulation $\mathcal{L}_c$, the depth gradient in pixel coordinate space $\left(\frac{\partial Z}{\partial u}, \frac{\partial Z}{\partial v}\right)$ depends on the absolute depth of the pixel as shown in equation 7, 8. So, our formulation takes into account the scale of separation between points over which the depth gradient is assumed to be linear. Furthermore, $\left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}\right) = \left(\frac{-n_x}{n_z}, \frac{-n_y}{n_z}\right)$ doesn't depend on the absolute depth value and hence only provides information about the relative depths of the pixels, where as $\left(\frac{\partial Z}{\partial u}, \frac{\partial Z}{\partial v}\right)$ depends on the absolute depth of the pixel locality too.

5. Semantic class specific evaluation on ScanNet

We quantify the performance of our methods on planar and textureless surfaces by evaluating on semantic classes on ScanNet test images. Specifically we use the eigen13 classes [1] and report the depth estimation metrics of our methods against DPSNet. We present the other frequently occuring classes not presented in the paper here in Table 1. We show that our methods perform well on all semantic categories and quantitatively show the improvement on planar and textureless surfaces as well which are usually found on walls, floors and ceiling.

6. KITTI 2015 Benchmark

We try to evaluate our method on the KITTI 2015 stereo benchmark. We pre-train our network on the Scene Flow datasets and finetune it on KITTI 2015 train data. We also pre-train GANet-NNet (defined in 4.2 in main paper) on Scene Flow datasets. For GANet, we use the pretrained models the authors provide. We test the performance of these pretrained models first on the KITTI train data without training on it and the report the EPE and 3 pixel error rate in Table 2. We then proceed to train on the KITTI 2015 train data and provide the results of the benchmark in Table 3

We observe that the pretrained models generalize better than other methods on KITTI 2015. We obtain significant improvement over DPSNet on the KITTI 2015 test set by adding normal supervision. The KITTI 2015 dataset contains only 200 training images with sparse ground truths with the sparsity increasing as we move to the background. Our ground truth normals are generated using a least squares optimization on the ground truth depths. Sparsity in the ground truth depths makes the generation of very accurate ground truth normals difficult. We see this as a significant problem and affects our performance on KITTI 2015. Despite this problem, GANet-NNet performs better than GANet on the foreground regions.

7. More Qualitative Results

We present more qualitative results on depth map estimation in Figure 2. The examples depict various situations

Label	Method	Abs Rel(↓)	Abs diff(↓)	Sq Rel(↓)	RMSE (↓)
Bed	DPSNet	0.1291	0.1572	0.050	0.1986
	Ours	0.1142	0.1449	0.0405	0.1830
	Ours-$\mathcal{L}_c$	**0.1049**	**0.1347**	**0.0345**	**0.1665**
Books	DPSNet	0.1087	0.2281	0.0733	0.2527
	Ours	0.0970	0.2176	0.0650	0.2404
	Ours-$\mathcal{L}_c$	**0.0942**	**0.2139**	**0.0628**	**0.2334**
Ceiling	DPSNet	0.1693	0.3429	0.1029	0.3895
	Ours	0.1496	0.3189	0.0840	0.3528
	Ours-$\mathcal{L}_c$	**0.1360**	**0.2244**	**0.0643**	**0.2900**
Chair	DPSNet	0.1602	0.2469	0.0836	0.3187
	Ours	0.1417	0.2351	0.0697	0.3050
	Ours-$\mathcal{L}_c$	**0.1360**	**0.2244**	**0.0643**	**0.2900**
Floor	DPSNet	0.1116	0.2472	0.0777	0.2973
	Ours	0.1092	0.2242	0.0509	0.2642
	Ours-$\mathcal{L}_c$	**0.1037**	**0.2061**	**0.0474**	**0.2561**
Objects	DPSNet	0.1305	0.2375	0.0785	0.2934
	Ours	0.1165	0.2237	0.0661	0.2771
	Ours-$\mathcal{L}_c$	**0.1095**	**0.2113**	**0.0589**	**0.2587**
Picture	DPSNet	0.1160	0.2991	0.0949	0.3249
	Ours	0.1110	0.2913	0.0912	0.3167
	Ours-$\mathcal{L}_c$	**0.1017**	**0.2724**	**0.0808**	**0.2923**
Table	DPSNet	0.1374	0.2211	0.0745	0.2808
	Ours	0.1238	0.2116	0.0646	0.2694
	Ours-$\mathcal{L}_c$	**0.1164**	**0.2014**	**0.0590**	**0.2545**
Wall	DPSNet	0.1340	0.2968	0.0871	0.3599
	Ours	0.1255	0.2835	0.0799	0.3436
	Ours-$\mathcal{L}_c$	**0.1173**	**0.2690**	**0.0721**	**0.3215**
Window	DPSNet	0.1559	0.3836	0.1353	0.4384
	Ours	0.1468	0.3605	0.1111	0.4163
	Ours-$\mathcal{L}_c$	**0.1373**	**0.3385**	**0.1079**	**0.3848**

Table 1. Semantic class specific evaluation on ScanNet. "DPSNet" corresponds to the predictions from DPSNet. "Ours" corresponds to our predictions before refinement by the consistency module. "Ours-$\mathcal{L}_c$" refers to our final predictions

Method	EPE(↓)	3-pixel error rate(↓)
GANet-deep	1.66	10.5
GANet-NNet	1.64	9.7
Ours	**1.64**	**8.2**

Table 2. Evaluation of Scene Flow pretrained models on KITTI2015. For all the metrics, lower the better.

like planar surfaces, reflective surfaces, planar-textureless surfaces and in general overall quality of the prediction. The red boxes on the images illustrate these regions. Our method produces more accurate depth maps when compared to the previous state-of-the-art.

Method	fg-noc(↓)	both-noc(↓)	fg-all(↓)	both-all(↓)
DPSNet	6.08	4.00	7.58	4.77
GANet-deep	3.11	**1.63**	3.46	**1.81**
GANet-NNet	**3.04**	1.70	**3.34**	1.91
Ours	4.06	2.08	4.41	2.27

Table 3. Comparative evaluation of our model on KITTI 2015 dataset. For all the metrics, lower the better. **fg:** Foreground, **both:** Foreground and Background, **noc:** Non occluded Pixels, **all:** All Pixels

References

[1] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Conference Track Proceedings*, 2013. 2

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2432–2443, 2017. 1

[3] David F Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3392–3399, 2013. 1

[4] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 1, 2

[5] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018. 1

[6] Runze Zhang, Shiwei Li, Tian Fang, Siyu Zhu, and Long Quan. Joint camera clustering and surface segmentation for large-scale multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2084–2092, 2015. 1