Deep Surface Normal Estimation on the 2-Sphere with Confidence Guided Semantic Attention

Quewei Li[1][0000−0003−3828−6756], Jie Guo[1], Yang Fei[1][0000−0001−7874−9175], Qinyu Tang[1][0000−0002−0130−7448], Wenxiu Sun[2], Jin Zeng[2], and Yanwen Guo[1]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2] SenseTime Research, China
{queweili, yangf, tangqinyu}@smail.nju.edu.cn,
{guojie, ywguo}@nju.edu.cn, {irene.wenxiu.sun, jzeng2010}@gmail.com

1 Introduction

In this supplementary material, we first illustrate the detailed implementation of the CGSA module. Then, we evidence the training stability and easy convergence for surface normal estimation on the 2-sphere, as compared with that in the 3D Euclidean space. Additionally, we provide quantitative analysis on a small testing list for Matterport3D, which is selected by Zeng *et al.* [8], to further validate the superior performance of our model. Finally, we present more visual comparisons with both the state-of-the-art methods and the different variants of our proposed method.

2 Implementation of the CGSA Module

The key point of the CGSA module is to update each low confidence depth feature patch with the most similar high confidence depth feature patch as described in Algorithm 1.

To be specific, we utilize the attention map $\mathcal{M}$ to divide $\mathcal{F}_c$ into M_c and $\bar{M}_c \in \mathbb{R}^{E\times H\times W}$ in line $2 \sim 4$, where E, H, W correspond to the channel number, height and width, respectively. $\otimes$ denotes the element-wise multiplication. ω is a pre-defined large constant that ensures patches in the attention regions will not be mistaken as reference patches. Then M_c and $\bar{M}_c$ are reshaped as $M_c \in \mathbb{R}^{HW\times E}$ and $\bar{M}_c \in \mathbb{R}^{E\times HW}$ in line $5 \sim 6$ for matrix multiplication. We compute the pair-wise similarities, *i.e.*, $\mathcal{S} \in \mathbb{R}^{HW\times HW}$ between M_c and $\bar{M}_c$ with the squared L_2 distance defined as $\|M_c - \bar{M}_c\|_2^2 = M_c^2 + \bar{M}_c^2 - 2M_c\bar{M}_c$ in line $7 \sim 10$. After that, we obtain the mapping function $\mathbf{\Psi} \in \mathbb{R}^{HW}$ via finding the index of the minimum value in each row of $\mathcal{S}$ in line 11. We apply $\mathbf{\Psi}$ to $\bar{\mathcal{F}}_d \in \mathbb{R}^{E\times HW}$ to get the reference depth feature map $\mathcal{F}_d^* \in \mathbb{R}^{E\times HW}$ in line $12 \sim 13$, where the i^{th} column of $\mathcal{F}_d^*$ is equal to the $\mathbf{\Psi}(i)^{th}$ column of $\bar{\mathcal{F}}_d$.

Algorithm 1: Confidence guided semantic attention

Input: the color feature map $\mathcal{F}_c$, the depth feature map $\mathcal{F}_d$ and the down-sampled confidence map $\hat{C}$

Output: the re-weighted depth feature map $\mathcal{F}'_d$

1 $E, H, W \leftarrow \mathcal{F}_c$.shape;
2 $\mathcal{M} \leftarrow \textbf{Float}(\hat{C} < 1)$;
3 $M_c \leftarrow \mathcal{F}_c \otimes \mathcal{M}$;
4 $\bar{M}_c \leftarrow \mathcal{F}_c \otimes (1 - \mathcal{M}) + \omega\mathcal{M}$;
5 $M_c \leftarrow M_c$.reshape(HW, E);
6 $\bar{M}_c \leftarrow \bar{M}_c$.reshape$(E, HW)$;
7 $X\_Y \leftarrow \textbf{MatrixMultiplication}(M_c, \bar{M}_c)$;
8 $X \leftarrow \textbf{Sum}(M_c \otimes M_c, \text{dim} = 1)$.repeat$(1, HW)$;
9 $Y \leftarrow \textbf{Sum}(\bar{M}_c \otimes \bar{M}_c, \text{dim} = 0)$.repeat$(HW, 1)$;
10 $\mathcal{S} \leftarrow (X + Y - 2X\_Y)$;
11 $\mathbf{\Psi} \leftarrow \textbf{Argmin}(\mathcal{S}, \text{dim} = 1)$;
12 $\bar{\mathcal{F}}_d \leftarrow \mathcal{F}_d$.reshape$(E, HW)$;
13 $\mathcal{F}^*_d \leftarrow \mathbf{\Psi}(\bar{\mathcal{F}}_d)$;
14 $\mathcal{F}^*_d \leftarrow \mathcal{F}^*_d$.reshape$(E, H, W)$;
15 $\mathcal{F}'_d \leftarrow \hat{C} \otimes \mathcal{F}_d + (1 - \hat{C}) \otimes \mathcal{F}^*_d$;

Finally, $\mathcal{F}^*_d$ are reshaped as $\mathcal{F}^*_d \in \mathbb{R}^{E \times H \times W}$ and the CGSA module outputs the re-weighted depth feature map $\mathcal{F}'_d$ with the following scheme in line $14 \sim 15$:

$$\mathcal{F}'_d = \hat{C} \otimes \mathcal{F}_d + (1 - \hat{C}) \otimes \mathcal{F}^*_d. \tag{1}$$

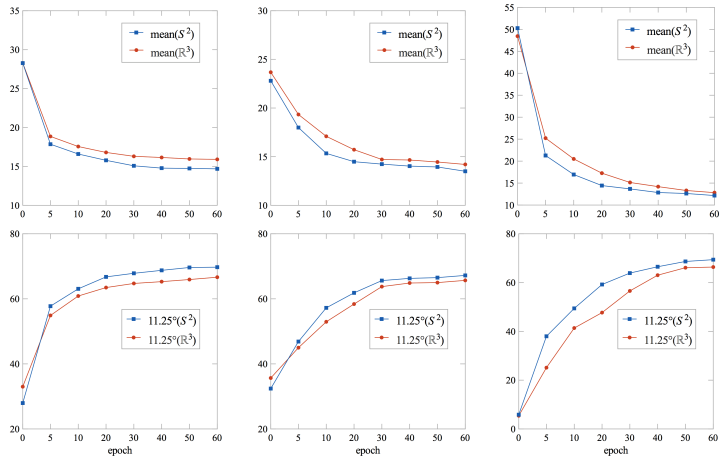
The algorithm can be easily converted into a batch version that deals with several depth feature maps within a batch group simultaneously. Since we convert the searching task into tasks of matrix multiplication and finding the index of the minimum value in each row of $\mathcal{S}$, the CGSA module is very efficient. Note that we detach $\mathcal{M}$ and $\mathbf{\Psi}$ during gradients backwards propagation. Therefore, the whole process of the CGSA module is differentiable.

3 Training Stability: 2-Sphere vs. 3D Euclidean Space

In this section, we conduct several experiments to verify the stable training and easy convergence of predicting surface normals on the 2-sphere, as compared with that in the 3D Euclidean space.

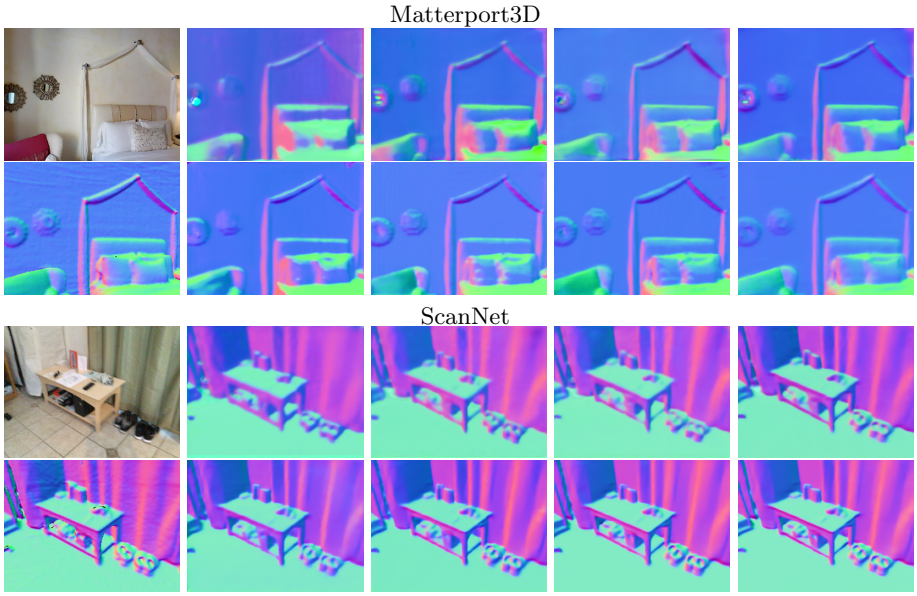
Since our training process carries on S^2, we convert the learned θ and ϕ back to the 3D Euclidean space after prediction for evaluation. The reverting formulas are:

$$\begin{cases} x = \cos\theta\cos\phi \\ y = \sin\theta\cos\phi \\ z = \sin\phi \end{cases} \tag{2}$$



(a) Matterport3D (b) ScanNet (c) NYUD-v2

Fig. 1. Evaluation metrics convergence between $\mathbb{R}^3$ and S^2 on Matterport3D, ScanNet and NYUD-v2 datasets



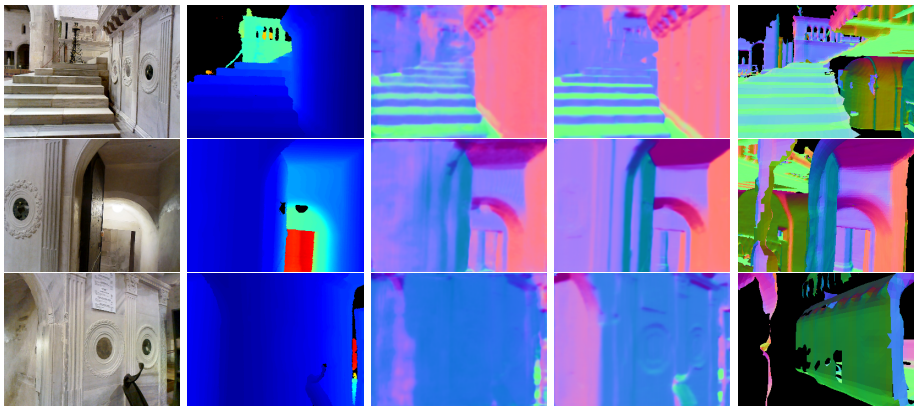
(a) GT (b) 5 (c) 15 (d) 35 (e) 60

Fig. 2. Visual quality comparisons between $\mathbb{R}^3$ (the first and third rows) and S^2 (the second and fourth rows) after 5, 15, 35 and 60 epochs on Matterport3D and ScanNet datasets. The ground truths are shown in the leftmost column

We compare surface normal estimation between $\mathbb{R}^3$ and S^2 with the evaluation metrics of the mean of angle error (mean) and angle difference less than $t_n = 11.25°$ after 1, 5, 10, 20, 30, 40, 50 and 60 epochs on Matterport3D [1], ScanNet [2] and NYUD-v2 [5] datasets. Moreover, we visualize the estimated results after 5, 15, 35, 60 epochs to better understand the convergence. As seen, both quantitative comparisons in Fig. 1 and visual comparisons in Fig. 2 reveal that the model trained on the 2-sphere enjoys a much faster convergence and achieves better performance when both models are converged.

4 Quantitative Analysis on the Selected Matterport3D Dataset

The ground-truth data provided by Zhang *et al.* [9] are generated with multi-view reconstruction. Zeng *et al.* [8] pointed out that some samples in the Matterport3D suffer from severe reconstruction error, which may lead to unreliable evaluation. Hence, they provided a new testing list removing samples with large errors (782 out of 12084). Fig. 3 provides some examples. To validate the robustness of our method, we also make quantitative comparisons on this dataset with the state-of-the-art methods. As shown in Table 1, our method still beats the state-of-the-arts on this testing dataset.



(a) Color (b) Raw depth (c) HFM-Net (d) Ours (e) Bad GT

Fig. 3. Several samples of ground-truth/target normals with obviously reconstruction errors

5 More Visual Quality Comparisons

In this section, we present more visual comparison results with the state-of-the-art surface normal estimation and depth recovery methods on Matterport3D,

Table 1. Performance of surface normal estimation on the new testing list of Matterport3D provided by Zeng *et al.* [8]

	Metrics	Zhang's	GeoNet	SharpNet	HFM-Net	Ours
Matter-port3D	mean	19.346	17.234	17.997	13.062	**12.336**
	median	9.762	8.744	10.061	5.270	**4.517**
	rmse	29.312	32.859	29.996	22.983	**22.934**
	11.25°	52.64	64.89	56.88	72.23	**73.73**
	22.5°	72.12	78.50	75.22	84.41	**85.37**
	30°	79.44	83.75	80.71	88.31	**89.11**

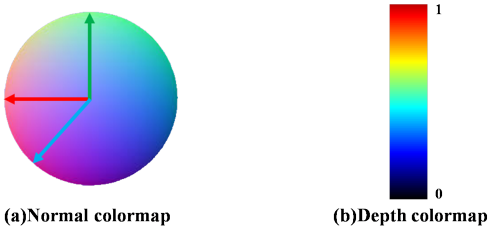


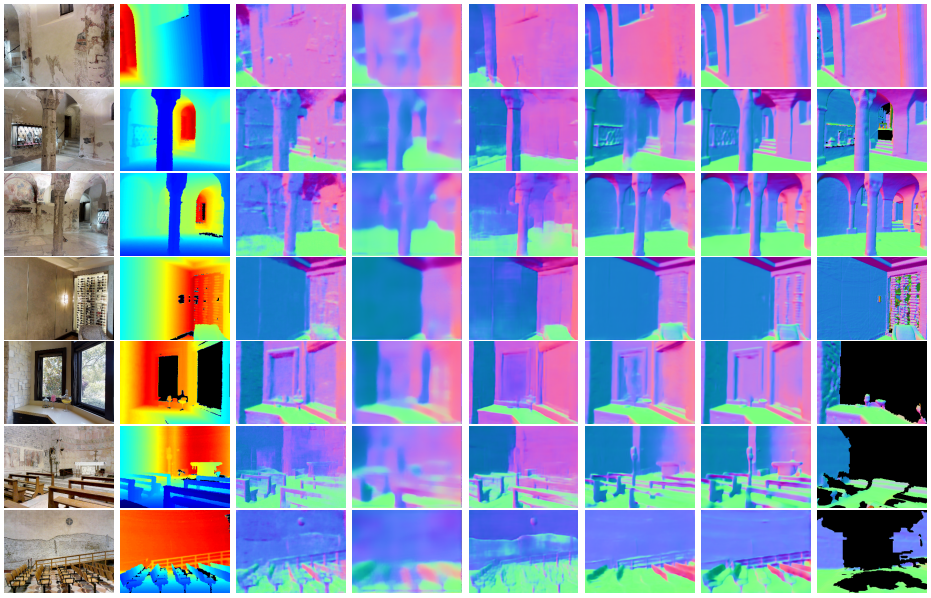
Fig. 4. Visualization of colormaps. The surface normal is visualized as red for left, green for up and blue for outward and the depth changes from blue to red with the depth values from 0 to 1

ScanNet and NYUD-v2 datasets. The methods in comparison are Zhang's network [10], GeoNet [6], SharpNet [7] and HFM-Net [8] for surface normal estimation and GeoNet [6], SharpNet [7], MonoD [3] and LabDEN [4] for depth recovery. Besides, more visual comparisons of different variants of the proposed method on the above three datasets are also provided in this section. The colormap for visualization is illustrated in Fig. 4. Specifically,

- Figures 5, 6 and 7 show the surface normal estimation comparisons on the three datasets, respectively.
- Figures 8, 9 and 10 show the depth recovery comparisons on the three datasets, respectively.
- Figure 11 compares all the different variants of our proposed method simultaneously.

References

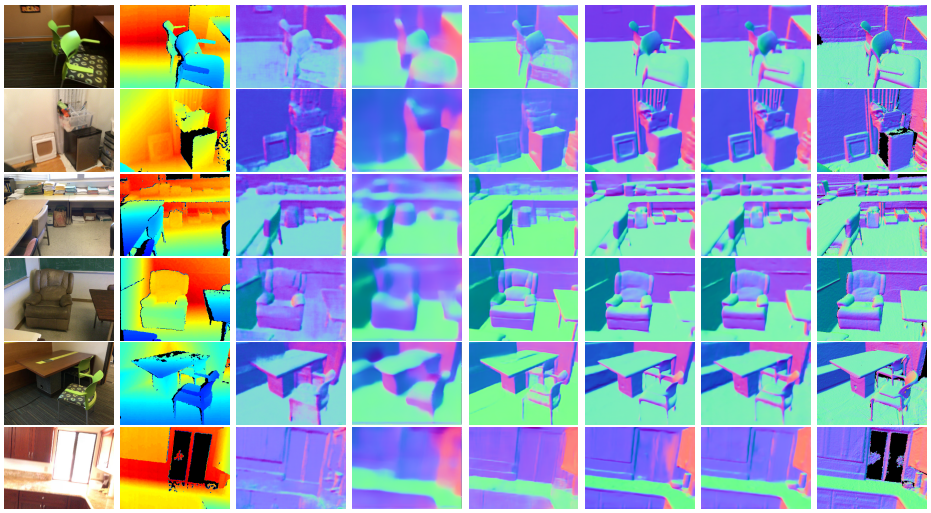
1. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV) (2017)
2. Dai, A., Nießner, M., Zollöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. ACM Transactions on Graphics 2017 (TOG) (2017)



(a)Color (b)Raw depth (c)Zhang's (d)GeoNet (e)SharpNet (f)HFM-Net (g)Ours (h)GT

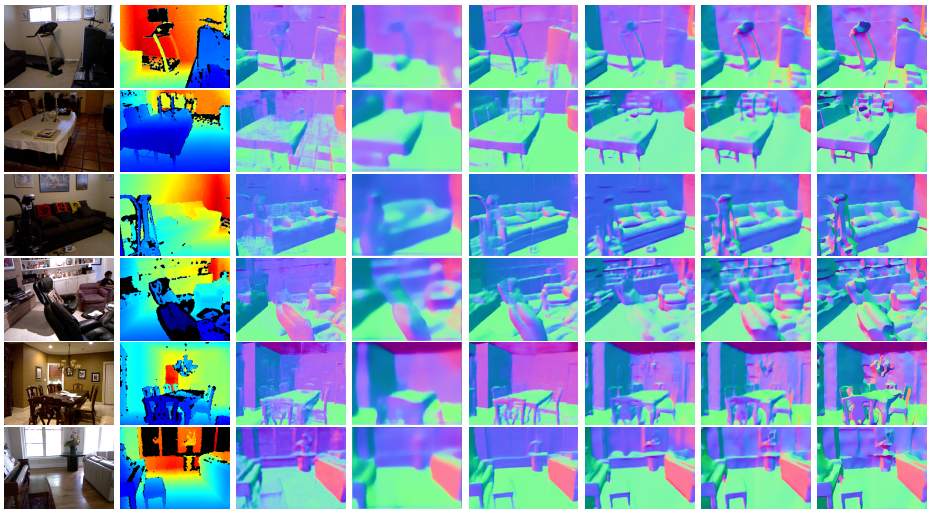
Fig. 5. Visual quality comparisons with the state-of-the-art surface normal estimation methods on Matterport3D

3. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
4. Jeon, J., Lee, S.: Reconstruction-based pairwise depth dataset for depth image enhancement using cnn. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 438–454. Springer International Publishing, Cham (2018)
5. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
6. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
7. Ramamonjisoa, M., Lepetit, V.: Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. The IEEE International Conference on Computer Vision (ICCV) Workshops (2019)
8. Zeng, J., Tong, Y., Huang, Y., Yan, Q., Sun, W., Chen, J., Wang, Y.: Deep surface normal estimation with hierarchical rgb-d fusion. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
9. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. Zhang, Y., Song, S., Yumer, E., Savva, M., Lee, J.Y., Jin, H., Funkhouser, T.: Physically-based rendering for indoor scene understanding using convolutional neural networks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)



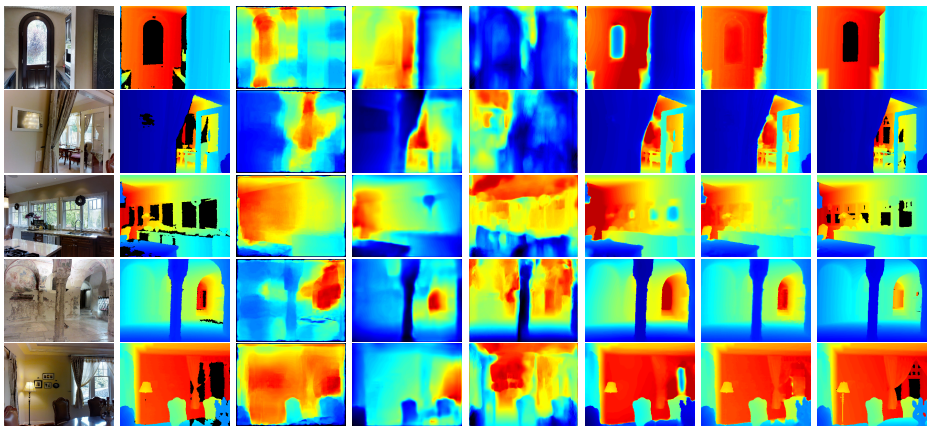
(a)Color (b)Raw depth (c)Zhang's (d)GeoNet (e)SharpNet (f)HFM-Net (g)Ours (h)GT

Fig. 6. Visual quality comparisons with the state-of-the-art surface normal estimation methods on ScanNet



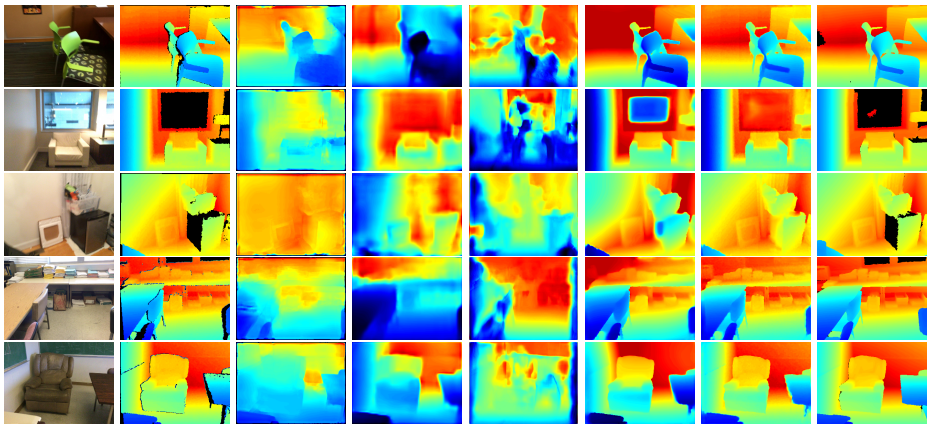
(a)Color (b)Raw depth (c)Zhang's (d)GeoNet (e)SharpNet (f)HFM-Net (g)Ours (h)GT

Fig. 7. Visual quality comparisons with the state-of-the-art surface normal estimation methods on NYUD-v2



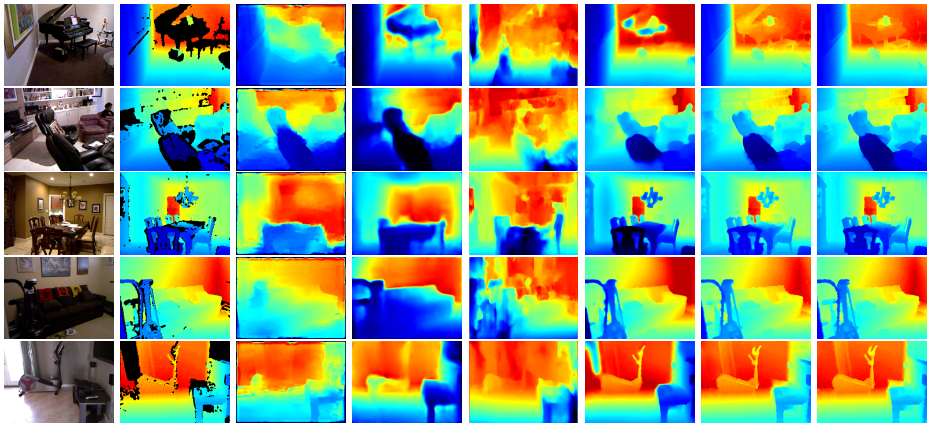
(a)Color (b)Raw depth (c)GeoNet (d)SharpNet (e)MonoD (f)LabDEN (g)Ours (h)GT

Fig. 8. Visual quality comparisons with the state-of-the-art depth recovery methods on Matterport3D



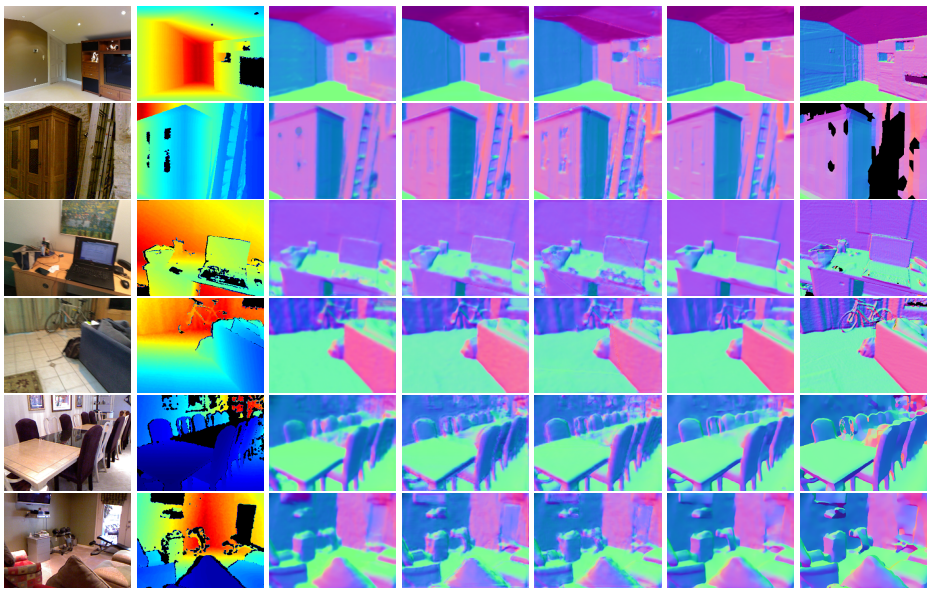
(a)Color (b)Raw depth (c)GeoNet (d)SharpNet (e)MonoD (f)LabDEN (g)Ours (h)GT

Fig. 9. Visual quality comparisons with the state-of-the-art depth recovery methods on ScanNet



(a)Color (b)Raw depth (c)GeoNet (d)SharpNet (e)MonoD (f)LabDEN (g)Ours (h)GT

Fig. 10. Visual quality comparisons with the state-of-the-art depth recovery methods on NYUD-v2



(a)Color (b)Raw depth (c)$\mathbb{R}^3$ (d)-CGSA (e)-MFF(f)Our complete (g)GT

Fig. 11. Visual quality comparisons on surface normal estimation of different variants