Supplementary Material for "Deep Surface Normal Estimation with Hierarchical RGB-D Fusion"

Jin Zeng[1] Yanfeng Tong[1,2*] Yunmu Huang[1*] Qiong Yan[1] Wenxiu Sun[1]
Jing Chen[2] Yongtian Wang[2]
[1]SenseTime Research [2]Beijing Institute of Technology
[1]{zengjin, tongyanfeng, huangyunmu, yanqiong, sunwenxiu}@sensetime.com
[2]{chen74jing29, wyt}@bit.edu.cn

1. Introduction

In this supplementary material, we discuss the data pruning process for Matterport3D [2] test set, then illustrate the network architecture details regarding the fusion network and confidence map estimation module. Additionally, we present more visual comparisons of surface normal estimation with competing schemes, with extra evaluation focusing on object details. Finally, more objective and visual results of variants of the proposed HFM-Net are presented for better understanding of the network.

2. Data Pruning

Since ground-truth normal data in the Matterport3D suffer from reconstruction noise, we remove the samples in Matterport3D testing set with large error using the following procedures:

- Compute the angle difference between ground-truth normal and results of two state-of-the-art methods (Skip-Net [1] and Zhang's method [7]);
- Remove the data item if the percentage of average normal angle difference within 45 degree is smaller than 90%.
- Manually check all erroneous samples to ensure that these samples suffers from incorrect multi-view reconstruction, *e.g.*, due to outdoor scenes or mirror area.

Finally 782 out of 12084 test images are removed. Note that data pruning is not applied on Matterport3D training set since the effect of erroneous data is reduced with the hybrid loss, nor on ScanNet dataset where the size of erroneous area is acceptable.

*indicates equal contribution.

3. HFM-Net Architecture

Details of the HFM-Net architecture are demonstrated in Fig. 2. Convolution and deconvolution layers are denoted as $Conv(in, out, k, s)$ and $Deconv(in, out, k, s)$, where in and out are input and output channel numbers, k is the kernel size and s is the stride. Max-pooling and max-unpooling layers are denoted as $Maxpool(k, s)$ and $MaxUnpool(k, s)$, where k is the kernel size and s is the stride. Instance normalization is shortened as IN. LeakyRelu functions in confidence map estimation module are set with negative slope 0.01. $\bigoplus$ stands for concatenation, and $\bigotimes$ stands for element-wise multiplication. $Repeat\ layer$ is to increase the channel number of confidence map from one to that of the corresponding depth feature to facilitate subsequent element-wise multiplication.

Note that $MaxUnpool$ layers in RGB decoder share the pooling indices used in $MaxPool$ in RGB encoder.

3.1. Additional Layers for Multi-Scale Output

To compute hybrid loss on multi-scale output in Eq. (4) in the paper, for each scale, we additionally add one convolution layer taking the feature output from the deconvolution layer before $MaxUnpool$ in RGB branch as the input, and outputs a 3-channel output using kernel size 3 and stride 1. In this way, the multi-scale outputs are generated to enable hybrid loss computation. These layers are removed in the testing mode.

4. Main Results

In this section, more visual comparison results with competing normal estimation schemes on Matterport3D [2] and ScanNet [3] are presented in Fig. 3, 4, 5, and 6. The colormap for visualization is illustrated in Fig. 1.

While RGB-based methods, *i.e.*, Skip-Net [1] and Zhang's algorithm [8], generate overall reasonable results, fine details are missing and the edges are blurred out. This

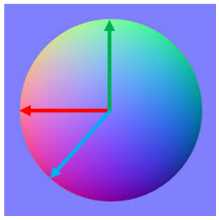


Figure 1: Surface normal visualization colormap. Red for left, green for up and blue for outward.

is because RGB-based methods suppress the sensitivity to sharp changes or fine textures to avoid mistakes at these regions, but in turn blur the edges in the final results. For results generated from depth with pre-processing, *i.e.*, inpainting with Levin's algorithm [5] and depth completion (DC) [7], the inpainted area is not as accurate as that predicted from RGB-D fusion methods. For results generated from alternative RGB-D fusion schemes, *i.e.*, incorporating depth input into RGB branch with methods in GeoNet (GeoNet-D) [6] or GFMM [4], the depth input is converted to normal map and then used to refine normal prediction in RGB branch, thus the results of GeoNet-D and GFMM are similar. Since the depth information is not fully utilized, the results are still blurry though additional details are introduced from depth input. On the other hand, HFM-Net provides more visually appealing results, with smooth surfaces and sharp details.

In addition, we visualize the normal results in 3D, in Fig. 7, 8, 9 and 10. We first convert the ground-truth depth map to the world coordinates with the provided intrinsic camera parameters to obtain corresponding point cloud, which is then rendered with normal estimation as texture. As shown in the figures, HFM-Net provides more accurate result in inpainted area compared to depth based methods, and preserves more details and sharp edges compared to RGB based and RGB-D methods.

Moreover, to stress the importance of fine details around objects, we additionally evaluate the performance in the object regions for ScanNet dataset [3] by computing pixels belonging to particular objects. Specifically, semantic segmentation results provided in ScanNet dataset are used to mask objects of interest, *i.e.*, bed, sofa, chair, and quantitative results of different approaches are shown in Table 1. As can be seen, HFM-Net outperforms the competing schemes by a large margin, which is consistent with the visual results in Fig. 3, 4, 5, 6. For example, by comparing with the state-of-the-art DC, we achieve **5.03-13.50%** increase in the terms of the fraction of pixels with angle error within 11.25°, indicating a better detail preservation.

Table 1: Performance of surface normal prediction for local object layout on ScanNet dataset. The best results are highlighted in boldface.

Objects	Metrics	RGB-based		Depth-based		RGBD-based		Preposed
		Skip-Net [1]	Zhang's [8]	Levin's [5]	DC [7]	GeoNet-D [6]	GFMM [4]	HFM-Net
Bed	mean	28.918	28.045	20.621	17.854	21.866	21.044	**14.288**
	median	25.381	23.777	12.991	11.673	18.604	17.853	**9.740**
	11.25°	17.19	23.42	52.79	55.26	26.68	31.93	**60.29**
	22.5°	45.05	49.73	71.79	75.96	63.42	63.66	**82.12**
	30°	60.30	62.53	77.93	82.37	75.75	76.62	**88.32**
Sofa	mean	25.801	25.31	22.391	19.056	22.479	21.527	**13.325**
	median	21.153	20.800	15.027	13.400	18.864	17.853	**8.737**
	11.25°	24.95	26.29	50.97	52.77	28.28	30.98	**66.27**
	22.5°	55.68	56.31	68.95	75.11	63.45	66.54	**84.77**
	30°	68.65	69.38	74.88	81.66	77.21	79.57	**89.53**
Chair	mean	34.209	32.601	34.950	27.640	28.958	28.268	**22.009**
	median	29.555	27.528	26.453	18.073	23.705	22.937	**13.840**
	11.25°	14.43	17.88	33.33	40.91	19.62	21.66	**48.75**
	22.5°	39.00	42.97	48.42	60.88	49.56	51.70	**69.10**
	30°	52.37	55.93	55.56	68.59	64.21	65.80	**76.28**

5. Ablation Study

In this section, we present a complete comparison with variants of the HFM-Net for understanding the role of each component in the network. We evaluate two categories of HFM-Net variants, one to validate fusion scheme, and the other to test loss function. Apart from that, we examine the effectiveness of confidence map, and compare the result of single branch and the complete HFM-Net.

5.1. Fusion Schemes

In the first category, we test on different fusion schemes, *i.e.*, early fusion (Early-F) and late fusion (Late-F), and keep masking and loss function unchanged if needed in the fusion. Multi-scale output is not involved in Early-F and Late-F, thus L_1 loss is used for training.

Visual comparison with Early-F and Late-F is shown in Fig. 11, where HF has smoother transition along the boundary of the depth holes. For example in the second row of Fig. 11, the boundary of the chair back marked with orange rectangle is smoother in the HFM-Net result.

Additionally, in category "Fusion(binary mask+hybrid)" of Table 2, HF outperforms Early-F and Late-F in all metrics. For example, for results of ScanNet dataset, HF outperforms Early-F and Late-F by 4.25 and 9.41%, suggesting a better detail preservation.

5.2. Loss Function

In the second category, we test on different loss functions. Besides to the comparison between L_2 loss and hybrid loss on HF with confidence map, we additionally compare L_2 and L_1 loss on the model used in Zhang's scheme [8] with RGB input. As shown in Fig. 12, hybrid/L_1 loss generates sharper edges and suppress errors in planar area. This is consistent with the results in Table 2, where "RGB L_1" and "HF hybrid" outperforms "RGB L_2" and "HF L_2" respectively, especially in terms of percentage of angle er-

Table 2: Performance of different variants of HFM-Net with hybrid loss to evaluate: hierarchical fusion scheme (HF) compared with early fusion (Early-F) and late fusion (Late-F), usage of confidence map compared with binary mask, hybrid loss compared with L_2 loss. The best results are highlighted in boldface.

Dataset	Metrics	Variants of HFM-Net						
		Fusion (binary mask+hybrid)			Loss (HF with Map)			
		Early-F	Late-F	HF	RGB L_2	RGB L_1	HF L_2	HF hybrid
Matterport3D	mean	13.968	13.645	13.437	19.346	18.913	13.688	**13.062**
	median	6.855	6.567	6.507	12.070	10.751	7.235	**6.090**
	11.25°	71.93	70.79	70.98	52.64	56.60	69.21	**72.23**
	22.5°	83.54	83.68	83.96	72.12	73.51	83.45	**84.41**
	30°	87.44	87.75	88.05	79.44	79.82	87.94	**88.31**
ScanNet	mean	16.045	17.425	14.696	23.306	22.575	14.946	**14.590**
	median	8.949	10.277	7.545	15.950	13.633	8.322	**7.468**
	11.25°	61.17	56.01	65.42	40.43	47.06	62.87	**65.65**
	22.5°	79.32	76.93	81.10	63.08	66.93	80.12	**81.21**
	30°	84.87	83.26	86.11	71.88	74.06	85.72	**86.21**

Table 3: Performance of single branch network and HFM-Net. The best results are highlighted in boldface.

Dataset	Metrics	RGB branch	Depth branch	HMF-Net
Matterport3D	mean	19.346	15.161	**13.062**
	median	12.070	8.148	**6.090**
	11.25	52.64	66.12	**72.23**
	22.5	72.12	81.12	**84.41**
	30	79.44	85.90	**88.31**
ScanNet	mean	23.306	16.689	**14.590**
	median	15.95	9.688	**7.468**
	11.25	40.43	58.55	**65.65**
	22.5	63.08	78.43	**81.21**
	30	71.88	84.31	**86.21**

ror under 11.25° because the performance improvement focuses on detail refinement.

5.3. Confidence Map

Furthermore, to evaluate the effectiveness of confidence map, we compare results of HF with binary mask (from the first category) and with confidence map (from the second category) and show in Fig. 13, which are denoted as HF(Mask) and HF(Map) respectively. With confidence map, the transition along depth holes are more accurate, which is further validated in Table 2 where "HF hybrid" in the second category provides better results than "HF" in the first category.

5.4. Single Branch Output

We also compare the result of using the single RGB branch, the single depth branch and the whole HFM-Net, and show in Table 3. Single RGB branch is the same as Zhang's [7]. It can be concluded from the result that the HFM-Net successfully integrates the features of two branches and achieves higher performance than a single branch.

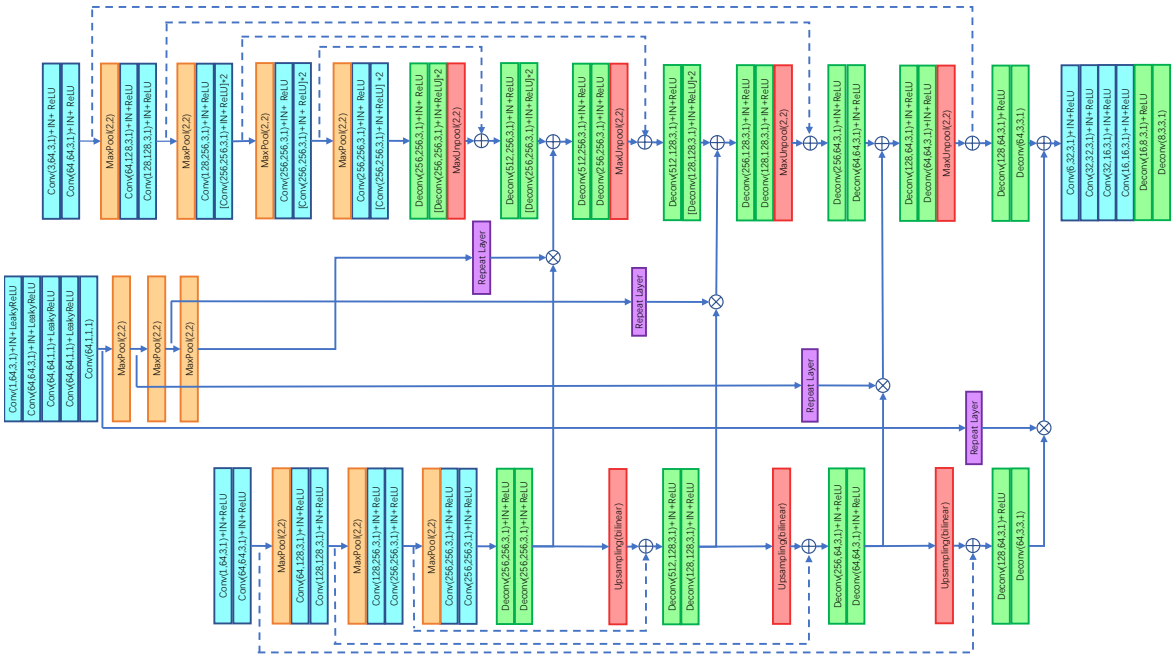
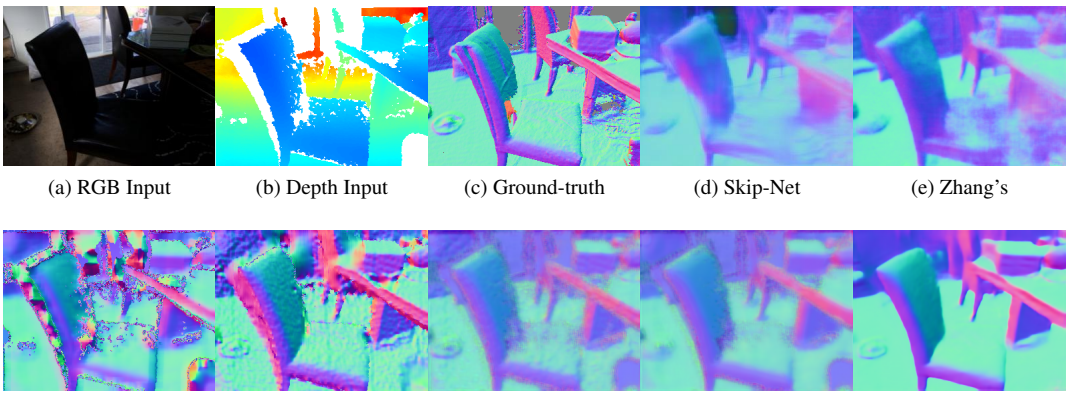


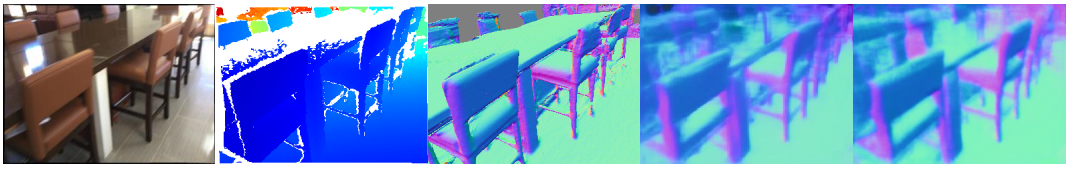
Figure 2: Details of HFM-Net architecture



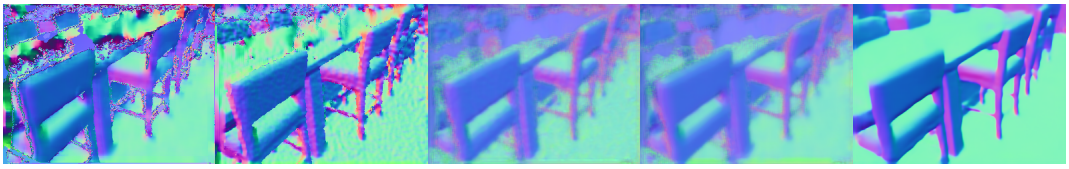
(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's

(f) Levin's (g) DC (h) GeoNet-D (i) GFMM (j) HFM-Net

Figure 3: Surface normal estimation with different schemes, tested on ScanNet dataset.



(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's



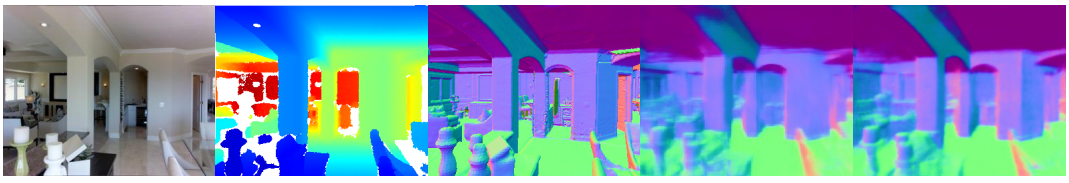
(f) Levin's (g) DC (h) GeoNet-D (i) GFMM (j) HFM-Net

Figure 4: Surface normal estimation with different schemes, tested on ScanNet dataset.

(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's

(f) Levin's (g) DC (h) GeoNet-D (i) GFMM (j) HFM-Net

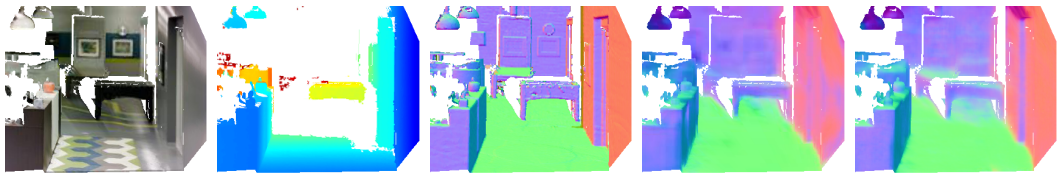
Figure 5: Surface normal estimation with different schemes, tested on Matterport3D dataset.



(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's

(f) Levin's (g) DC (h) GeoNet-D (i) GFMM (j) HFM-Net

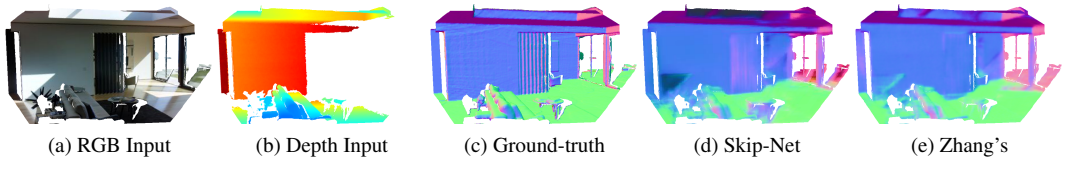
Figure 6: Surface normal estimation with different schemes, tested on Matterport3D dataset.



(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's

(f) Levin's (g) DC-inpainting (h) GeoNet-D (i) GFMM (j) HFM-Net

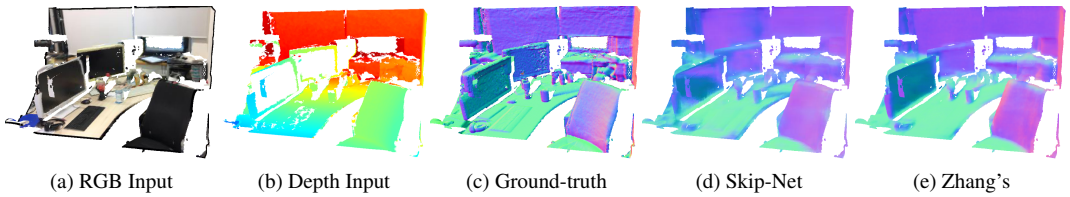
Figure 7: 3D visualization of normal estimation with different schemes, tested on Matterport3D dataset.



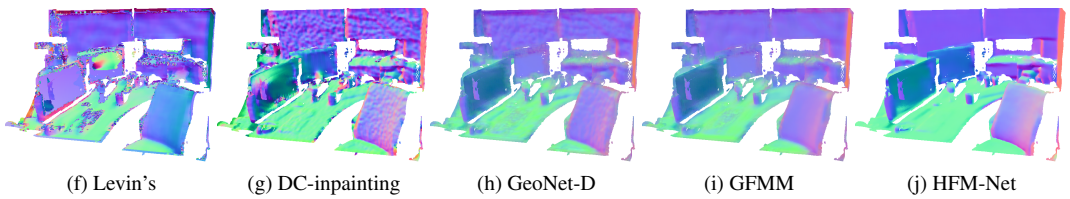
(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's

(f) Levin's (g) DC-inpainting (h) GeoNet-D (i) GFMM (j) HFM-Net

Figure 8: 3D visualization of normal estimation with different schemes, tested on Matterport3D dataset.

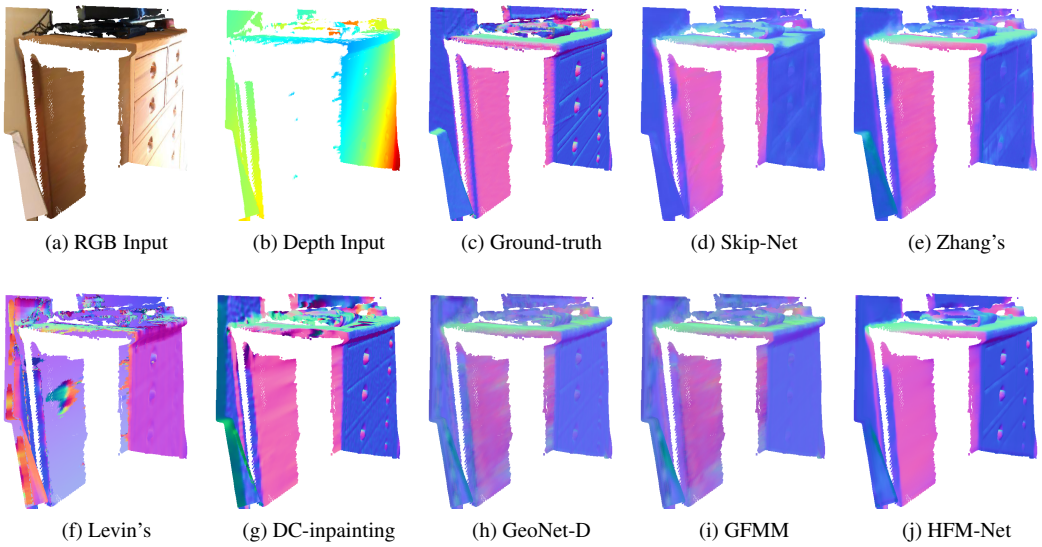


(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's



(f) Levin's (g) DC-inpainting (h) GeoNet-D (i) GFMM (j) HFM-Net

Figure 9: 3D visualization of normal estimation with different schemes, tested on ScanNet dataset.



(a) RGB Input (b) Depth Input (c) Ground-truth (d) Skip-Net (e) Zhang's

(f) Levin's (g) DC-inpainting (h) GeoNet-D (i) GFMM (j) HFM-Net

Figure 10: 3D visualization of normal estimation with different schemes, tested on ScanNet dataset.

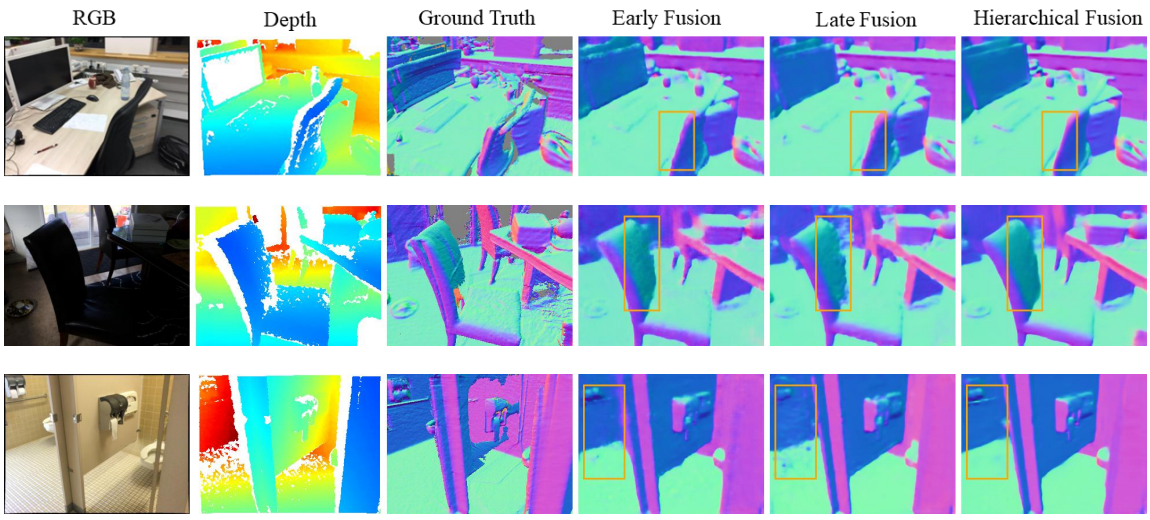


Figure 11: Surface normal estimation with different fusion schemes.

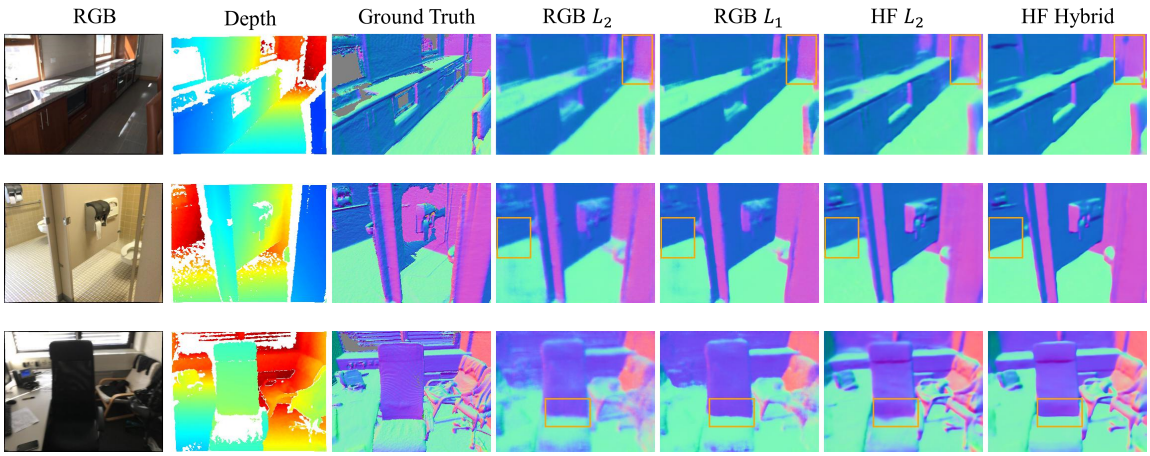


Figure 12: Surface normal estimation with different loss functions.

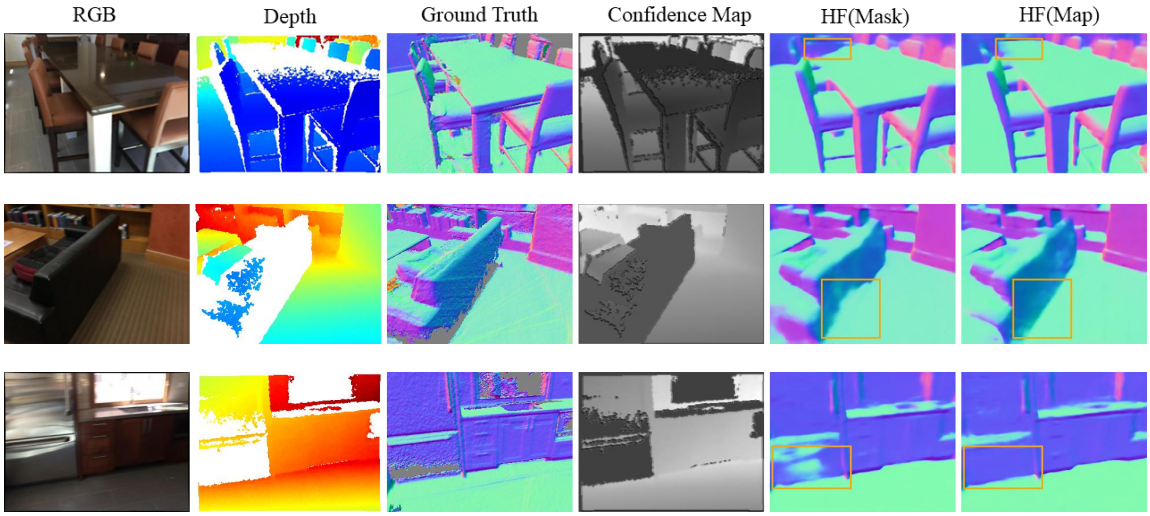


Figure 13: Surface normal estimation with binary mask and confidence map.

References

[1] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5965–5974, 2016. 1, 2

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1

[3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[4] X. Gong, J. Liu, W. Zhou, and J. Liu. Guided depth enhancement via a fast marching method . *Image & Vision Computing*, 31(10):695–703, 2013. 2

[5] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *ACM transactions on graphics (tog)*, volume 23, pages 689–694. ACM, 2004. 2

[6] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. 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. 2

[7] Y. Zhang and T. Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–185, 2018. 1, 2, 3

[8] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5057–5065. IEEE, 2017. 1, 2