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

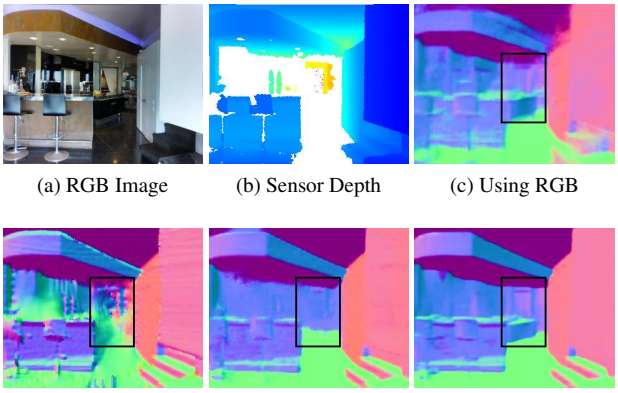
Abstract

The growing availability of commodity RGB-D cameras has boosted the applications in the field of scene understanding. However, as a fundamental scene understanding task, surface normal estimation from RGB-D data lacks thorough investigation. In this paper, a hierarchical fusion network with adaptive feature re-weighting is proposed for surface normal estimation from a single RGB-D image. Specifically, the features from color image and depth are successively integrated at multiple scales to ensure global surface smoothness while preserving visually salient details. Meanwhile, the depth features are re-weighted with a confidence map estimated from depth before merging into the color branch to avoid artifacts caused by input depth corruption. Additionally, a hybrid multi-scale loss function is designed to learn accurate normal estimation given noisy ground-truth dataset. Extensive experimental results validate the effectiveness of the fusion strategy and the loss design, outperforming state-of-the-art normal estimation schemes.

1. Introduction

Per-pixel surface normal estimation has been extensively studied in the recent years. Previous works on normal estimation mostly assume single RGB image input [8, 26, 1, 33], providing satisfying results in most cases despite loss of shape features and erroneous results at the highlight or dark areas, as shown in Fig. 1(c).

RGB-D cameras are now commercially available, leading to a great performance enhancement in the applications of scene understanding, *e.g.*, semantic segmentation [27, 5, 23], object detection [11, 20], 3D reconstruction [15, 18, 12], *etc*. With the depth given by sensors, normal can be easily calculated via a least square optimization [21, 9] as used in the widely used NYUv2 dataset [22],



(a) RGB Image (b) Sensor Depth (c) Using RGB
(d) Using Depth (e) Early Fusion (f) Hierarchical Fusion

Figure 1. Example in Matterport3D dataset. (a) RGB input; (b) depth input; normal estimation with (c) single RGB [33], (d) depth inpainting [9], (e) RGB-D early fusion [32], (f) proposed hierarchical RGB-D fusion.

but the quality of the normal suffers from the corruption in depth, *e.g.*, sensor noise along object edges or missing pixels due to glossy, black, transparent, and distant surfaces [24], as shown in Fig. 1(d).

This motivates us to combine the advantages of color and depth inputs while compensating for the deficiency of each other in the task of normal estimation. Specifically, the RGB information is utilized to fill the missing pixels in depth; meanwhile the depth clue is merged into RGB results to enhance sharp edges and correct erroneous estimation, resulting in a complete normal map with fine details. However, research on combining RGB and depth for normal estimation has not been extensively studied. To the best of our knowledge, the only work considering RGB-D input for normal estimation adopts *early fusion*, *i.e.*, using depth as an additional channel to the RGB input, leading to little performance improvement compared with the methods using the RGB input only [32]. The lack of proper network design for combining the geometric information in depth and color image is an impediment to fully take advantage

*indicates equal contribution.

of the depth sensor.

Different from previous works on normal estimation with RGB-D using early fusion [32], we propose to merge the features from RGB and depth branches at multiple scales at the decoder side in a hierarchical manner, in order to guarantee both global surface smoothness and local sharp features in the fusion results. Additionally, a pixel-wise confidence map is estimated from the depth input for re-weighting depth features before merging into RGB branch, so as to reduce artifact from depth with a smaller confidence on missing pixels and those along the object edges. An example is shown in Fig. 1, where the proposed scheme outperforms state-of-the-art RGB-based, depth-based, RGBD-based methods.

Apart from the lack of RGB-D fusion schemes, the shortage of datasets providing sensor depth and ground-truth depth pairs is another obstacle for RGB-D normal estimation since the performance of DNN approaches is affected by the dataset quality [19, 30]. The widely used training datasets for normal estimation, *e.g.*, NYUv2 [22], do not provide complete ground-truth normal for the captured RGB-D images since it is directly computed from the captured depth after inpainting [14]. If trained on NYUv2, the network is up to approximate an inpainting algorithm.

Instead we use Matterport3D [2] and ScanNet [6] datasets with RGB-D captured by camera and ground-truth normal obtained via multiview reconstruction provided by [32]. Nevertheless, the ground-truth is not perfect due to the multiview reconstruction error, especially at object edges which is crucially for visual evaluation. To overcome the artifact in the ground-truth, we propose a hybrid multi-scale loss function based on the noise statistics in the ground-truth normal map, using L_1 loss at the large resolution to obtain sharper results, and L_2 loss at small resolution to ensure coarse scale accuracy.

In summary, the main contributions of our work are:

- By incorporating RGB and depth inputs via the proposed hierarchical fusion scheme, the two inputs are able to complement each other in the normal estimation, refine details with depth, and fill the missing depth pixels with color;
- With the confidence map for depth feature re-weighting, the effect of artifacts in the depth features is reduced;
- A hybrid multi-scale loss function is designed by analyzing the noise statistics in the ground-truth, providing sharp results with high fidelity despite the imperfect ground-truth.

Comparison with the state-of-the-art approaches and extensive ablation study validates the design of network structure and loss function. The paper is organized as follows. Related works are discussed in Section 2, and Section 3 provides a detailed discussion of the proposed method. Ablation study and comparison with state-of-the-art methods are demonstrated in Section 4 and the work is concluded in Section 5.

2. Related Work

2.1. Surface Normal Estimation

RGB-based Previous works mostly used a single RGB image as input. Eigen *et al*. [8] designed a three-scale convolution network architecture that produced a coarse global prediction with full image first and then refined it with local finer-scale network. Wang *et al*. [28] proposed a network structure that integrated different geometric information like local, global, and vanishing point information to predict the surface normal. More recently, Bansal *et al*. [1] proposed a skip-connected structure to concatenate the CNN response at different scales to capture corresponding details at each scale, and Zhang *et al*. [33] adopted a U-Net structure and achieved state-of-the-art performance.

Due to the difficulty in extracting geometric information and texture interference from the RGB input, the details of predictions are poor, with wrong results in the area of insufficient lighting or high lighting.

Depth-based Surface normal can be inferred from depth with geometric method, which depends on the neighboring pixels' relative depth geometrically [32]. However, the depth camera used in common datasets, *e.g.*, NYUv2 [22], Matterport3D [2], ScanNet [6] often fails to sense the depth on glossy, bright, transparent and faraway surfaces [32, 29], resulting in holes and corruptions in the obtained depth images. To overcome missing pixels in normal map inferred from depth, some works proposed to inpaint depth images using RGB images [7, 10, 16, 25, 31]. Silberman *et al*. [22] used optimization-based method [14] to fill the holes in depth maps. Zhang *et al*. [32] used a convolutional network to predict pixel-wise surface normal with a single RGB image, then used the predicted normal to fill holes in raw depth.

Nevertheless, depth inpainting cannot handle large holes in depth; also, the noise in depth will undermine depth-based normal estimation performance.

Normal-depth consistency based There is a strong geometric correlation between the depth and the surface normal. Normal can be calculated from the depth of neighboring pixels, and depth can be refined with normal variation. For example, Wang *et al*. [26] proposed a four-stream convolutional neural network to detect planar regions, then used a dense conditional random field to smooth results based on depth and surface normal correlation in planar region and planar boundary respectively. Chen *et al*. [3] es-

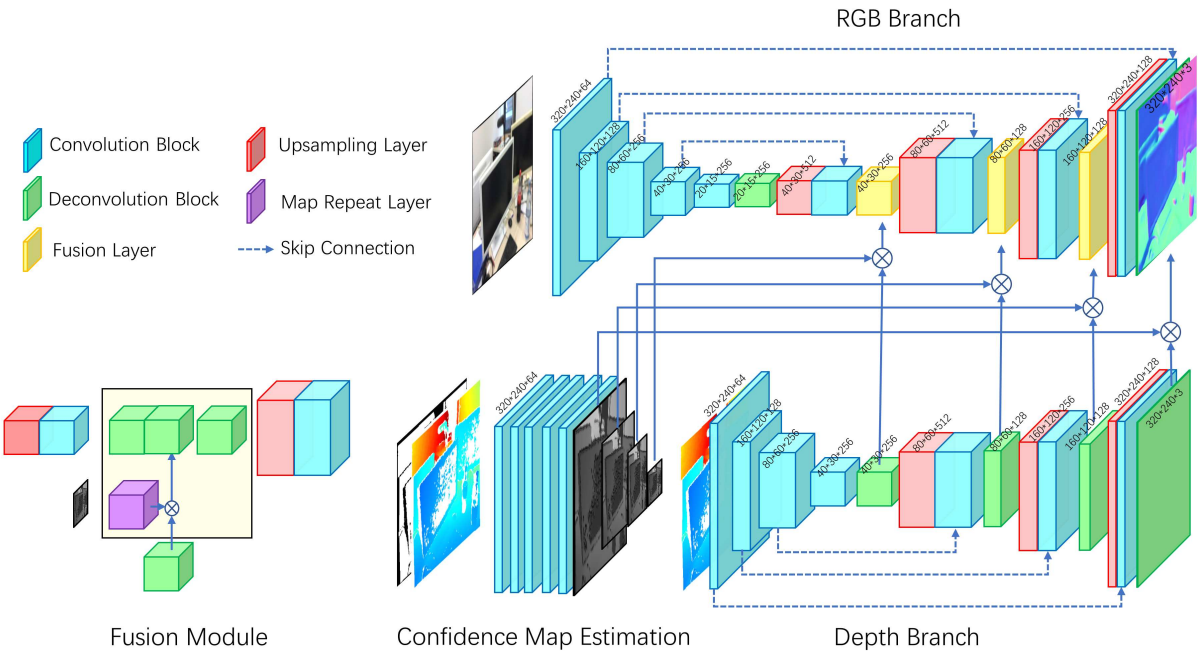


Figure 2. Proposed hierarchical RGB-D fusion composed of RGB branch at the upper side, depth branch at the lower-right side, confidence map module at the lower-left side. The fusion module is abstracted as a fusion layer in the fusion network and illustrated at the lower-left side. An input with size 320×240 is used for demonstration.

tablished a new dataset, and proposed two loss functions to measure the consistency between predicted normal and depth label for depth and normal prediction. Qi *et al*. [21] proposed to predict initial depth and surface normal using color image, then cross-refine each other using geometric consistency.

These methods provide different schemes to promote geometric consistency between normal and depth, but rely on a single RGB input and do not consider noise from depth sensors.

RGB-D based The RGB-D based normal estimation has not been extensively studied in previous works. Normal estimation with RGB-D input has been briefly discussed in [32] where an early fusion was adopted, reported to be almost the same as using RGB input. However the method is not properly designed and the conclusion is not comprehensive. Although 3D reconstruction based methods like [18] can be used in normal estimation, a series of RGB-D images is required for those methods, which is beyond the scope of this paper. The lack of design in RGB-D fusion for surface normal estimation motivates our work.

2.2. RGB-D Fusion Schemes

Despite the lack of study in RGB-D based normal estimation, RGB-D fusion scheme has been explored for other tasks, among which semantic segmentation is the most extensively studied one, *e.g*., early fusion using RGB-D as a four-channel input [8], late fusion [4], depth-aware convolution [27], or using 3D point cloud format [20].

The difference from those works is that they do not require per-pixel accuracy as much as normal prediction, *i.e.*, the label interior of one object is constant, but for normal estimation, correct prediction at each pixel is required, and the most significant difficulty lies in accurate sharp details. Therefore, we adopt hierarchical fusion with confidence map re-weighting to enhance edge preservation in the fusion result without bringing artifacts in depth.

3. Method

As illustrated in Fig. 2, the hierarchical RGB-D fusion network is composed of three modules: RGB branch, depth branch, and confidence map estimation. In this section, we introduce the pipeline for the hierarchical fusion of RGB and depth branches with the fusion module at different scales, and confidence map estimation used inside the fusion module for depth conditioning, after which the hybrid loss function design is detailed. A detailed architecture of the deep network is provided in the supplementary.

3.1. Hierarchical RGB-D Fusion

Given color image $\mathcal{I}_c$ and sensor depth $\mathcal{I}_d$, we are aimed as estimating surface normal map $\mathcal{I}_n$ by minimizing its distance from the ground-truth normal $\mathcal{I}_n^{(\mathrm{gt})}$, *i.e.*,

$$\min_{\theta} \quad L(\mathcal{I}_n^{(\mathrm{gt})}, f_\theta(\mathcal{I}_c, \mathcal{I}_d)), \tag{1}$$

where f_θ denotes the fusion network function to generate normal estimation $\mathcal{I}_n$ parameterized by the parameters θ, which are end-to-end trained via back propagation. A hierarchical fusion scheme is adopted to merge depth branch into RGB branch for both overall surface orientation rectification and visually salient feature enhancement.

3.1.1 Network Design

First, in the RGB branch where the input is the color image $\mathcal{I}_c$, we adopt a similar network structure as used in [33], where a fully convolutional network (FCN) [17] is built with VGG-16 back-bone as illustrated in the RGB branch in Fig. 2. Specifically, the encoder is the same as VGG-16 except that in the last two convolution blocks of the encoder, *i.e.*, conv4 and conv5, the channel number is reduced from 512 to 256 to remove redundant model parameters. The encoder is accompanied with a symmetric decoder, and equipped with skip-connections and shared pooling masks for learning local image features.

Meanwhile, $\mathcal{I}_d$ is fed into the depth branch to extract geometric features with a similar network structure as the RGB branch, except that the last convolution block in the RGB encoder is removed to give a simplified model.

The fusion takes place at the decoder side. As shown in Fig. 2, the depth features (colored in green) at each scale in the decoder are passed into the fusion module and re-weighted with the confidence map (colored in purple) down-sampled and repeated to the same resolution as the depth feature. Then the re-weighted depth features are concatenated with the color features with the same resolution and passed through a deconvolution layer to give the fusion output features (colored in yellow). Consequently, the fusion module (denoted as FM for short) at scale l is given as,

$$\mathrm{FM}(\mathcal{F}_c^l, \mathcal{F}_d^l | \mathcal{C}^l) = \mathrm{deconv}(\mathcal{F}_c^l \oplus (\mathcal{F}_d^l \odot \mathcal{C}^l)), \tag{2}$$

where $\mathcal{F}_c^l$, $\mathcal{F}_d^l$ are the features from RGB and depth branches at scale l, and $\mathcal{C}^l$ is the confidence map for depth conditioning. $\odot$ denotes element-wise multiplication and $\oplus$ denotes the concatenation operation. The concatenation result after deconvolution layer gives the fusion output. The fusion is implemented at four scales, where the last scale output gives the final normal estimation. The confidence map estimation is addressed later in Section 3.2.

3.1.2 Comparison with Existing RGB-D Fusion Schemes

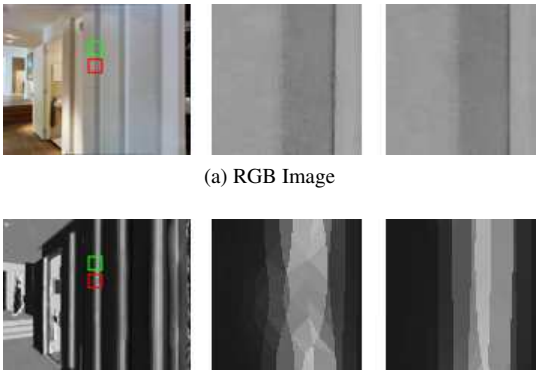
Existing RGB-D fusion schemes mostly adopt single-scale fusion. [32] fused RGB-D at the input, *i.e.*, using depth as an additional channel along with RGB. However, RGB and depth are from different domains and cannot be properly handled using the same encoder as a four-channel input. For example, we adopt the same network structure as in [33], composed of VGG-16 encoder and a symmetric decoder with skip-connection, and use a RGB-D four-channel input instead of a single RGB to generate the normal as shown in Fig. 7(d). The output normal does not exhibit global smoothness, especially in area where depth pixels are missing. This is because a CNN network is incapable of handling different domains information from RGB and depth without prior knowledge about depth artifact.

Late fusion with probability map for RGB and depth is adopted in [4] for segmentation, and here we generalize the network structure for normal estimation, by replacing the probability map with a binary mask indicating whether the depth pixel is available or not, giving the result in Fig. 7(e). The role of binary mask we use is consistent with that of the probability map in [4] which indicates how much the source is trustworthy. Similar to early fusion, the result of late fusion has noticeable artifacts along the depth holes indicating the fusion is not smooth.

In light of this, single-scale fusion is not efficient for fusing RGB and depth when RGB and depth contain different noise. RGB is sensitive to lighting conditions while depth is corrupted at object edges and distant surfaces, indicating that the output from RGB and depth can be inconsistent. If depth is integrated into RGB in a single scale, the fusion is hard to eliminate the difference between two sources and give a smooth result. This motivates us to merge depth features into RGB branch at four different scales in a hierarchical manner. In this way, the features from two branches are successively merged, where the global surface orientation error would be corrected at small resolution features, while detail refinement would take place at the final scale. As shown in Fig. 7, the result of the proposed hierarchical fusion gives smoother result with detail well preserved.

3.2. Confidence Map Estimation

While hierarchical fusion improves normal estimation over existing fusion schemes, further examination at pixels around depth holes shows that the transition is not smooth as shown in Fig. 8(e) where the right side of the table has erroneous prediction close to depth hold boundary. This indicates that a binary masking is not sufficient for depth conditioning, and a more adaptive re-weighting would be more favorable. Therefore, a light-weight network for depth confidence map is designed as follows.



(a) RGB Image

(b) Corresponding ground-truth normal

Figure 3. Enlarged patches from input image and ground-truth normal map in horizontal direction. Upper row: input image, patch in red rectangle, patch in green rectangle. Bottom row: ground-truth normal map, patch in red rectangle, patch in green rectangle.

Depth along with a binary mask indicating missing pixels in depth are fed into a convolutional network with five layers as shown in Fig. 2, where the first two layers are with 3×3 kernel size and the following three layers are with 1×1 kernels. In this way, the receptive field is small enough to restrict local adaption to depth variation. Then the confidence map is down-sampled using shared pooling mask with depth branch and passed into the fusion module to facilitate fusion operation as described in Eq. 2. By comparing Fig. 8(e) and (f), the confidence map leads to a more accurate fusion result, correcting the error at the right side of the table.

To understand the role of the confidence map, we show the confidence map in Fig. 8(d). The edge pixels are with the smallest confidence value indicating a high likelihood of outlier or noise, while the hole area is with a small yet non-zero value, suggesting that to enable smooth transition, information in depth holes can be passed into the merge result as long as RGB features take the dominant role.

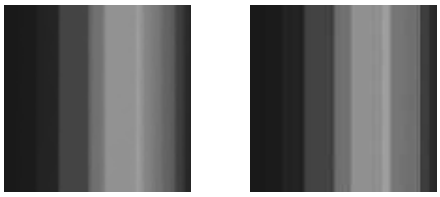
3.3. Hybrid Loss

As mentioned in Section 1, we use Matterport3D and ScanNet datasets for training and testing because RGB-D data captured by camera and ground-truth normal pairs are provided. However, the ground-truth normal suffers from multiview reconstruction errors as shown in Fig. 3(b) where the normal map is piece-wise constant inside the mesh triangular and the edge does not align with the RGB input. Given noisy ground-truth like this, improper handling of loss function during training will lead to deficient performance. The reason is as follows.

Given the similar inputs in green and red rectangular in Fig. 3(a), the output would be similar. However, the corresponding ground-truth normal maps are different as shown Fig. 3(b), thus by minimizing the loss function, the network will learn an expectation of all pairs of input and ground-truth [13]:

$$\min_{\theta} \quad \mathbb{E}_{(\mathcal{I}_c,\mathcal{I}_d,\mathcal{I}_n^{(\mathrm{gt})})} L(\mathcal{I}_n^{(\mathrm{gt})}, f_\theta(\mathcal{I}_c,\mathcal{I}_d)). \tag{3}$$

For L_2 loss $L_2(\mathcal{I}_n^{(\mathrm{gt})},\mathcal{I}_n) = \|\mathcal{I}_n^{(\mathrm{gt})} - \mathcal{I}_n\|_2^2$, the minimization will lead to an arithmetic mean of the observations, while L_1 loss $L_1(\mathcal{I}_n^{(\mathrm{gt})},\mathcal{I}_n) = |\mathcal{I}_n^{(\mathrm{gt})} - \mathcal{I}_n|$ will lead to median of the observations.



(a) mean (b) median

Figure 4. Mean and median results from normal observations with the same RGB input.

To see which loss is more proper for the given dataset, we sample patches along the edge in Fig. 3 with same horizontal position as patches in the color rectangles, and compute the mean and median normal results of these sampled patches shown in Fig. 4 where both generate reasonable results though median result has sharper edges than mean result, indicating that L_1 loss will generate a more visually appealing result with sharp details.

In this work, we adopt hybrid multi-scale loss function:

$$\begin{aligned} L(\mathcal{I}_n^{(\mathrm{gt})},\mathcal{I}_n) = &\sum_{l=1,2} w_l L_2(\mathcal{I}_n^{(\mathrm{gt})}(l),\mathcal{I}_n(l)) \\ &+ \sum_{l=3,4} w_l L_1(\mathcal{I}_n^{(\mathrm{gt})}(l),\mathcal{I}_n(l)), \end{aligned} \tag{4}$$

where $l = 1,2,3,4$ denotes the scales from small to large, and w_l is the weight for loss at different scales and is set to be $[0.2, 0.4, 0.8, 1.0]$. L_1 loss is used for large scale outputs for detail enhancement, while L_2 loss is used for coarse scale outputs for overall accuracy. Using hybrid loss generates clean and visually better result than L_2 loss widely used for normal estimation [21, 33, 1] as shown in Fig. 7. The proposed method is named as *Hierarchical RGB-D Fusion with Confidence Map*, and referred to as *HFM-Net* for short.

4. Experiment

4.1. Implementation Details

Dataset We evaluate our approach on two datasets, Matterport3D [2] and ScanNet [6]. For the corresponding ground-truth normal data, we use the render normal provided by [32] which was generated with multiview reconstruction. Matterport3D is divided into 105432 images for

		RGB-based		Depth-based		RGBD-based		Ours
	Metrics	Skip-Net [1]	Zhang's [33]	Colorization [14]	DC [32]	GeoNet-D [21]	GFMM [10]	HFM-Net
Matter-port3D	mean	26.081	19.346	21.588	19.126	17.234	16.537	**13.062**
	median	19.089	12.070	12.079	9.563	8.744	8.028	**6.090**
	11.25°	31.76	52.64	58.07	61.48	64.89	65.3	**72.23**
	22.5°	57.61	72.12	69.59	74.08	78.5	79.94	**84.41**
	30°	67.60	79.44	75.00	79.22	83.75	84.16	**88.31**
Scan-Net	mean	26.174	23.306	33.071	30.652	23.289	21.174	**14.590**
	median	20.598	15.95	23.451	20.762	15.725	13.598	**7.468**
	11.25°	28.78	40.43	34.52	39.35	46.41	50.78	**65.65**
	22.5°	54.30	63.08	49.47	55.27	64.04	67.30	**81.21**
	30°	67.00	71.88	56.37	60.03	76.78	77.00	**86.21**
	runtime	2.501s	0.039s	0.156+0.9s	0.156+0.058s	0.156+0.041s	0.156+0.041s	0.085s

Table 1. Performance of surface normal prediction on Matterport3D and ScanNet dataset.

training and 11302 for testing; ScanNet is divided into 59743 for training and 7517 for testing with file lists provided in [32]. Since ground-truth normal data in the Matterport3D suffer from reconstruction noise, *e.g.*, in outdoor scenes or mirror area, we remove the samples in the testing dataset with large error so as to avoid unreliable evaluation. After data pruning, 6.47% (782 out of 12084) testing images are removed, leading to 11302 remaining. Details of data pruning can be found in the supplementary.

Training Details We use RMSprop optimizer with initial learning rate set to $1e^{-3}$ and decayed at epoch $[2, 4, 6, 9, 12]$ with decay rate 0.5. The model is trained from scratch without pretrained model for 15 epochs. We first use L_2 loss for all scales in the first 4 epochs and then change to hybrid loss defined in Eq. 4 to ensure stable training at the beginning. We implement with PyTorch on NVIDIA GeForce GTX Titan X GPU.

Evaluation Metrics The normal prediction performance is evaluated with five metrics. We compute the per-pixel angle distance between prediction and ground-truth, then compute mean and median for valid pixels with given ground-truth normal. In addition to mean and median, we also compute the fraction of pixels with angle difference with ground-truth less than t where t = 11.25°, 22.5°, and 30° as used in [9].

4.2. Main Results

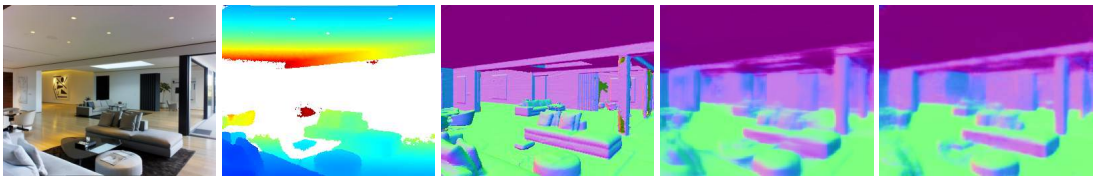
We compare our proposed HFM-Net with the state-of-the-art normal estimation methods, which are classified into three categories in accordance with Section 2, while normal-depth consistency based methods are adopted as alternatives for RGB-D fusion thus also put in the RGB-D category.

RGB-based methods include Skip-Net [1] and Zhang's algorithm [33]. Pretrained models on Matterport3D and ScanNet of Zhang's are provided in [32], and Skip-Net is fine-tuned for Matterport3D and ScanNet based on the pre-trained model on NYUv2 dataset using public available training code.

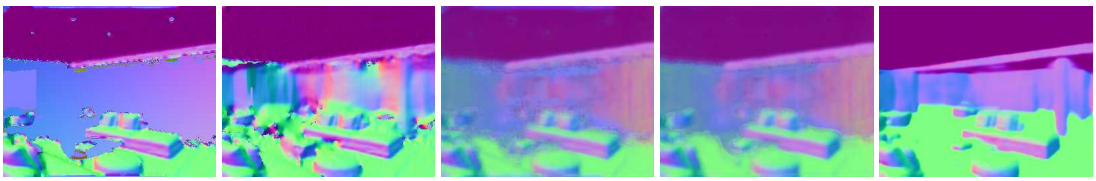
Depth-based Depth information is used to compute surface normal in existing works [22, 6, 2] based on geometric relation between depth and surface normal. Since the input depth is incomplete, we first implement depth inpainting before converting into normal map. Two algorithms are used to preprocess the input depth images: colorization algorithm in [14] as used in NYUv2 and the state-of-the-art depth completion (shortened as DC) [32]. After depth inpainting, we follow the same procedure in [21] to generate normal from depth.

RGBD-based For the RGB-D fusion methods, we adopt methods in GFMM [10] and the state-of-the-art GeoNet [21] to merge depth input into initial RGB-based normal output for refinement. Specifically, we choose Zhang's method [33] for initial normal estimation from RGB, and calculate a rough normal from raw depth image at the same time, then merge the two normal estimations using methods in GFMM [10] and GeoNet [21] to estimate the final surface normal map.

We test on two datasets respectively with the five metrics as shown in Table 1, where HFM-Net outperforms all the other schemes in different metrics. In terms of mean value, HFM-Net outperforms RGB-based methods by at least 6.284, and 6.064 over depth-inpainting based methods, and 3.475 over RGBD-based methods. Visual evaluation results are shown in Fig. 5 and Fig. 6. RGB-based methods miss details such as the sofa in Fig. 5 with blurry edges. Depth-based methods have serious errors at the depth hole regions and noticeable noise. Competing RGB-D fusion methods fail to generate accurate results at areas where depth is noisy or corrupted. On the contrary, our HFM-Net

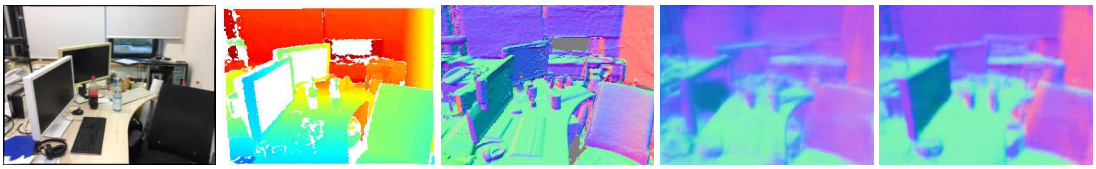


(a) RGB Image (b) Depth Image (c) Ground-truth (d) Skip-Net [1] (e) Zhang's [33]



(f) Colorization [14] (g) DC [32] (h) GeoNet-D [21] (i) GFMM [16] (j) HFM-Net

Figure 5. Surface normal estimation with different algorithms, test on Matterport3D dataset.



(a) RGB Image (b) Depth Image (c) Ground-truth (d) Skip-Net [1] (e) Zhang's [33]

(f) Colorization [14] (g) DC [32] (h) GeoNet-D [21] (i) GFMM [16] (j) HFM-Net

Figure 6. Surface normal estimation with different algorithms, test on ScanNet dataset.

is exhibiting nice normal prediction both at smooth planar areas and along sharp edges.

4.3. Ablation Study

For better understanding of how HFM-Net works, we investigate the effect of each component in the network with the following ablation study.

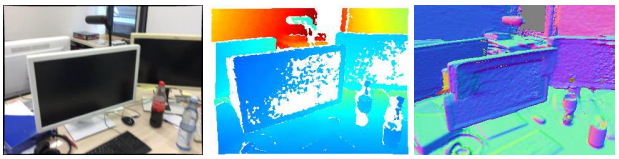
Hierarchical Fusion We compare hierarchical fusion (HF) with single-scale fusion including early fusion and late fusion as described Section 3, denoted as Early-F and Late-F in Table 2 respectively. The binary mask is used for Late-F and HF, and all are trained using hybrid loss if not specified. As can be seen from Table 2, Early-F and Late-F is less effective than HF+Mask+Hybrid, validating the use of HF. Furthermore, Fig. 7(d-f) show the difference between single-scale and hierarchical fusion. The hierarchical fusion provides more accurate results in a planar surface especially in depth hole areas marked in black rectangles.

Confidence Map We compare confidence map with binary mask. Fig. 8 shows the difference between fusion with confidence map and fusion with binary mask. Fusion with confidence map can reduce the negative effect of a depth hole during the fusion, and smooth the prediction around the boundary region of depth holes.

Hybrid Loss Apart from fusion method, different combinations of loss function are examined in the experiment. In comparison of hybrid loss, the confidence map is used in fusion. If the network use L_2 loss function in all layers, the prediction will tend to be blurry. On the other hand, a network with L_1 loss will tend to preserve more details. A hybrid loss function design, as described in Section 3.3 can generate results with both smooth surface and fine object details, as shown in the comparison in Fig. 7 (g-l).

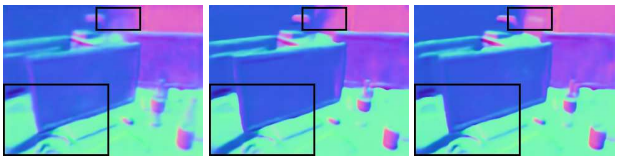
4.4. Model Complexity and Runtime

Table 1 reports the runtime of our method and other state-of-the-art methods. Skip-Net method uses the official evaluation code in MatCaffe. Colorization method uses the code provided in NYUv2 dataset. GeoNet-D is the GeoNet with RGBD input, and we implement it in PyTorch. The



(a) RGB Image (b) Sensor Depth (c) Ground-truth

(d) Early Fusion (e) Late Fusion (f) Hierarchical Fusion



(g) L2 Loss (h) L1 Loss (i) Hybrid loss

(j) L2 Loss Detail (k) L1 Loss Detail (l) Hybrid Loss Detail

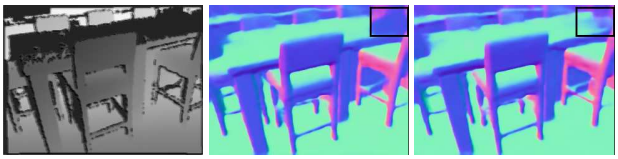
Figure 7. Surface normal estimation with different fusion schemes and different loss functions: (a) RGB input, (b) depth input, (c) ground truth, result of (d) early fusion, (e) late fusion, and (f) hierarchical fusion; result of using (g) L_2 loss, (h) L_1 loss, (i) proposed hybrid loss; (j-l) are the enlarged patches from (g-i). The hierarchical fusion produces a more accurate prediction in the area marked in black rectangles. The hybrid loss design preserves the advantages of both L_2 (smooth surface) and L_1 loss (local details), with sharper details and more accurate results in depth holes.

	Metrics	Early-F	Late-F	HF+Map +L2	HF+Mask +Hybrid	HF+Map +Hybrid
Matter-port3D	mean	13.968	13.645	13.688	13.437	**13.062**
	median	6.855	6.567	7.235	6.507	**6.090**
	11.25°	71.93	70.79	69.21	70.98	**72.23**
	22.5°	83.54	83.68	83.45	83.96	**84.41**
	30°	87.44	87.75	87.94	88.05	**88.31**
Scan-Net	mean	16.045	17.425	14.946	14.696	**14.590**
	median	8.949	10.277	8.322	7.545	**7.468**
	11.25°	61.17	56.01	62.87	65.42	**65.65**
	22.5°	79.32	76.93	80.12	81.10	**81.21**
	30°	84.87	83.26	85.72	86.11	**86.21**

Table 2. Evaluation of variants of the proposed HFM-Net on Matterport3D and ScanNet datasets.

consistency loss is added to GeoNet-D as a comparison scheme. The network forward runtime is averaged over Matterport3D test set with input images of size 320×256 on NVIDIA GeForce GTX TITAN X GPU. Apart from the time cost in neural network forward pass, the runtime of depth-based and RGBD-based methods also includes the time spent on geometric calculation. As in shown in Table 1, our method exceeds competing schemes in metric performance while taking a reasonably fast time.

(a) RGB Image (b) Sensor Depth (c) Ground-truth



(d) Confidence-map (e) HF (Mask) (f) HF (Map)

Figure 8. Surface normal estimation with different map/mask: (a) RGB input, (b) depth input, (c) ground truth, (d) confidence map, (e) hierarchical-fusion with mask, (f) hierarchical-fusion with map.

5. Conclusion

In this work, we propose a hierarchical fusion scheme to combine RGB-D features at multiple scales with a confidence map estimated from depth input for depth conditioning to facilitate feature fusion. Moreover, a hybrid loss function is designed to generate clean normal estimation even if the training targets suffer from reconstruction noise. Extensive experimental results demonstrate that our HFM-Net outperforms the state-of-the-art methods in providing more accurate surface normal prediction and sharper visually salient features. Ablation studies validate the superiority of the proposed hierarchical fusion scheme over single-scale fusion schemes in existing works, the effectiveness of confidence map in producing accurate estimation around missing pixels in depth input, and the advantage of the hybrid loss function in overcoming dataset deficiency.

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, 5, 6, 7

[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. 2, 5, 6

[3] W. Chen, D. Xiang, and J. Deng. Surface normals in the wild. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy*, pages 22–29, 2017. 2

[4] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang. Locality sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. 3, 4

[5] H. Chu, W.-C. M. K. Kundu, R. Urtasun, and S. Fidler. Surfconv: Bridging 3d and 2d convolution for rgbd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3002–3011, 2018. 1

[6] 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. 2, 5, 6

[7] H. C. Daniel, J. Kannala, L. Ladick, and J. Heikkil. *Depth Map Inpainting under a Second-Order Smoothness Prior*. Springer Berlin Heidelberg, 2013. 2

[8] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2650–2658, 2015. 1, 2, 3

[9] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3d primitives for single image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3392–3399, 2013. 1, 6

[10] 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, 6

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*, pages 345–360. Springer, 2014. 1

[12] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 1

[13] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2Noise: Learning image restoration without clean data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2965–2974, 2018. 5

[14] 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, 7

[15] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1886–1895, 2018. 1

[16] J. Liu, X. Gong, and J. Liu. Guided inpainting and filtering for kinect depth maps. In *International Conference on Pattern Recognition*, pages 2055–2058, 2012. 2, 7

[17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 4

[18] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1, 3

[19] J. Pang, W. Sun, C. Yang, J. Ren, R. Xiao, J. Zeng, and L. Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[20] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[21] 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. 1, 3, 5, 6, 7

[22] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 1, 2, 6

[23] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2530–2539, 2018. 1

[24] S. Su, F. Heide, G. Wetzstein, and W. Heidrich. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6383–6392, 2018. 1

[25] A. K. Thabet, J. Lahoud, D. Asmar, and B. Ghanem. 3d aware correction and completion of depth maps in piecewise planar scenes. In *Asian Conference on Computer Vision*, pages 226–241, 2014. 2

[26] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016. 1, 2

[27] W. Wang and U. Neumann. Depth-aware cnn for rgb-d segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2018. 1, 3

[28] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–547, 2015. 2

[29] J. Zeng, G. Cheung, M. Ng, J. Pang, and C. Yang. 3d point cloud denoising using graph laplacian regularization of a low dimensional manifold model. *arXiv preprint arXiv:1803.07252*, 2018. 2

[30] J. Zeng, J. Pang, W. Sun, G. Cheung, and R. Xiao. Deep graph laplacian regularization. *arXiv preprint arXiv:1807.11637*, 2018. 2

[31] H.-T. Zhang, J. Yu, and Z.-F. Wang. Probability contour guided depth map inpainting and superresolution using non-local total generalized variation. *Multimedia Tools and Applications*, 77(7):9003–9020, 2018. 2

[32] 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, 4, 5, 6, 7

[33] 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, 4, 5, 6, 7