Learning to Reconstruct and Understand Indoor Scenes From Sparse Views

Jingyu Yang, *Senior Member, IEEE,* Ji Xu, Kun Li, *Member, IEEE,* Yu-Kun Lai, *Member, IEEE,* Huanjing Yue, *Member, IEEE,* Jianzhi Lu, Hao Wu, and Yebin Liu, *Member, IEEE*

***Abstract*—This paper proposes a new method for simultaneous 3D reconstruction and semantic segmentation for indoor scenes. Unlike existing methods that require recording a video using a color camera and/or a depth camera, our method only needs a small number of (*e.g.*, 3∼5) color images from uncalibrated sparse views, which significantly simplifies data acquisition and broadens applicable scenarios. To achieve promising 3D reconstruction from sparse views with limited overlap, our method first recovers the depth map and semantic information for each view, and then fuses the depth maps into a 3D scene. To this end, we design an iterative deep architecture, named *IterNet*, to estimate the depth map and semantic segmentation alternately. To obtain accurate alignment between views with limited overlap, we further propose a joint global and local registration method to reconstruct a 3D scene with semantic information. We also make available a new indoor synthetic dataset, containing photorealistic high-resolution RGB images, accurate depth maps and pixel-level semantic labels for thousands of complex layouts. Experimental results on public datasets and our dataset demonstrate that our method achieves more accurate depth estimation, smaller semantic segmentation errors, and better 3D reconstruction results over state-of-the-art methods.**

***Index Terms*—3D reconstruction, deep learning, semantic segmentation, indoor scenes, sparse views**

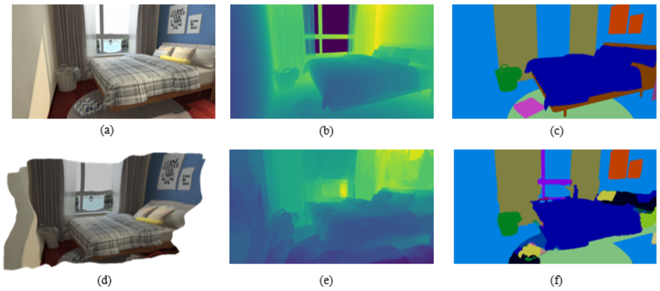


Figure 1. An example of 3D reconstruction by the proposed IterNet from five color images at sparse views: (a) one of the input RGB images, (b) ground-truth depth map, (c) ground-truth semantic segmentation, (d) our reconstructed 3D model using (e) the estimated depth map and (f) semantic labels. This example is part of test data.

I. Introduction

With the increasing demand for indoor navigation, home/office design, and augmented reality, indoor 3D reconstruction and understanding have become active topics in computer vision and graphics. Existing reconstruction methods can be broadly categorized into two groups. The first group scans indoor scenes with an integrated depth camera based on either time-of-flight (ToF) or structured light sensing. The pioneering KinectFusion [1] presents a detailed workflow using Kinect for indoor reconstruction. It was more recently extended by BundleFusion [2] which achieves state-of-the-art results in real-time 3D reconstruction. However, the quality of captured depth maps is affected by the limitations of sensing techniques, *e.g.*, limited sensing distances, low resolution, invalid measurements for specular areas, and/or significant amount of noise. On the contrary, RGB cameras provide much higher quality of color images, and are far more available particularly on smart phones. It is therefore interesting to study 3D scene reconstruction using a color camera, which however is challenging due to lack of depth information. Simultaneous localization and mapping (SLAM) [3] and structure from motion (SFM) [4] are two popular approaches for feature-based point cloud 3D reconstruction on-line and off-line, respectively. However, these feature-based methods require rich textures in the scene, and are therefore difficult to obtain dense point clouds. All the above methods require consecutive frame tracking or dense view capturing.

In this paper, we propose a new indoor-scene 3D reconstruction and semantic segmentation method using color images at several uncalibrated sparse views. The first challenge is to reconstruct a dense 3D scene from sparse views with limited overlap, which is practically degenerated into the task of monocular depth estimation. The second challenge, hence, is the non-rigid transformation between sparse views brought in by the inaccurate depth estimated from the single color image for each view. To address these problems, we design an iteratively optimized deep framework, named *IterNet*, for simultaneous depth map recovery and semantic segmentation for each view, where the two tasks help improve each other. To estimate complex transformations between sparse views, we further develop a joint global and local alignment method

This work was supported in part by the National Natural Science Foundation of China (61771339, 61672378 and 61531014), Tianjin Research Program of Application Foundation and Advanced Technology (18JCYBJC19200), and the National Key Research and Development Program of China (2018YFB2100500).

Jingyu Yang, Ji Xu, Hao Wu, and Huanjing Yue are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China.

Kun Li is with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, and the Key Research Center for Surface Monitoring and Analysis of Cultural Relics (SMARC), State Administration of Cultural Heritage, China.

Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, United Kingdom.

Jianzhi Lu is with the 3VJ Co., Ltd., Guangzhou 510000, China.

Yebin Liu is with the Department of Automation, Tsinghua University, Beijing 10084, China.

Corresponding author: Kun Li (Email: lik@tju.edu.cn)

to fuse estimated depths with the help of semantic information, which integrates geometry, photometry and semantic information in a coarse-to-fine manner.

Depth recovery and semantic segmentation from images are ill-posed, and it is essential to learn from high-quality training data. For indoor scene understanding, a number of datasets have been made publicly available. Real-world datasets, such as NYUv2 [5], SUN RGB-D [6], and ScanNet [7], need a lot of manual labor to annotate the labels and contain unavoidable noise in depths assumed as ground-truths, while synthetic datasets [8], [9] are difficult to generate photorealistic RGB images and usually have limited layouts and image resolution. To our best knowledge, no existing dataset can provide photorealistic RGB images, accurate depth maps, pixel-level semantic labels, and thousands of complex layouts at the same time. To address this, we build our IterNet RGB-D dataset with these features.

Experimental results on both public datasets and our dataset demonstrate that our method outperforms state-of-the-art methods on depth estimation, semantic segmentation, and sparse-view 3D reconstruction. Figure 1 gives an example of our IterNet RGB-D dataset and the reconstructed 3D model with estimated semantics using our IterNet. We will make the code and the dataset available online for research purposes.

In summary, our method is an integrated work that includes 1) a depth estimation method from a single color image; 2) a semantic segmentation method from a single view; and 3) a multi-view reconstruction method from sparse views. They jointly solve a challenging problem of 3D reconstruction and understanding from uncalibrated sparse views. As a result, our method is applicable to more scenarios than previous approaches that rely on textures and/or geometries of dense views, *e.g.*, reconstructing and understanding a room using only several photos captured by different users. Our main contributions are:

- We build the IterNet RGB-D dataset including photorealistic high-resolution RGB images, accurate depth maps, and pixel-level semantic labels for thousands of complex layouts. This is particularly useful for the training and evaluation of data-hungry learning-based approaches.
- We design a novel iterative learning-based method for joint depth estimation and semantic segmentation. Unlike previous learning-based methods that require two-way input (color+depth or color+semantic), our IterNet needs only a single color image for each view, and the depth-estimation branch and semantic segmentation branch are iteratively concatenated to help improve each other. The parallel context layer and atrous spatial pyramid pooling are incorporated to improve the LSTM (Long Short-Term Memory) module for better semantic segmentation. This architecture is not restricted to these tasks we address here, and is also applicable to other related tasks such as object/part parsing.
- We propose a joint global and local registration method to fuse different sparse perspectives. A seven-dimensional (7D) feature descriptor consisting of 3D location, photometric, and semantic information is designed for global registration. To improve the performance, a local alignment strategy is applied to semantic objects. This coarse-to-fine alignment is robust to sparse views and the errors of monocular depth estimation.

II. Related work

Indoor Datasets. Naseer *et al.* [12] gave a comprehensive overview of indoor scene understanding in 2.5D/3D. NYU-Depth dataset [5] captured by Microsoft Kinect is the first dataset of this category. SUN RGB-D dataset [6] captured by four different RGB-D sensors contains 10,335 indoor images with dense annotations. Armeni *et al.* [10] provided Building Parser dataset with instance level semantic and geometric annotations. Matterport3D [11] contains 10,800 panoramic images covering 360° viewpoints captured by a Matterport camera. ScanNet [7] is a 3D reconstruction dataset with 2.5 million frames obtained from 1,513 scans. The depth maps in these real-world datasets usually contain noise and missing areas and need a lot of manual effort for label annotation. Hence, synthetic datasets are constructed for easy generation and accurate ground truth. SUNCG [8] is a densely annotated large-scale indoor dataset, but the rendered RGB images are not photorealistic and RGB-D videos are not available. SceneNet RGB-D [9] provides pixel-level annotations and photorealistic RGB images, but the number of layouts is limited. Table I lists the main features of various publicly available 2.5/3D indoor datasets including our one. Our IterNet RGB-D dataset provides a total of 12,856 photorealistic images for thousands of layouts, and has higher image resolutions: 1280×960 or 1280×720, covering more indoor scenes. Moreover, our dataset provides absolute depth maps and pixel-level semantic segmentation that are more precise and accurate. Compared with other datasets, the indoor scenes covered by our dataset are more general and more complex.

Monocular Depth Estimation. In computer vision, monocular depth estimation has been a long-standing topic in last decades. Previous approaches mainly relied on hand-crafted features [13], statistical priors [14] or graphical models [15]. With the development of deep learning, more recent approaches were based on Convolutional Neural Networks (CNNs). For instance, Eigen *et al.* [16] proposed a multi-scale CNN for depth estimation, which achieves promising results. Considering the correlation between tasks, Wang *et al.* [17] introduced a CNN for joint depth estimation and semantic segmentation. Xu *et al.* [18] proposed a multi-task approach for depth estimation via cross-modal interactions. Recently, the attention mechanism has been extensively applied to various learning tasks, and Xu *et al.* [19] proposed a structured attention scheme for depth estimation to fuse features of different scales. Xu *et al.* [20] utilized a continuous Conditional Random Field (CRF) to combine multi-scale features. Our method also uses CRFs as basic building blocks for depth inference, but further integrates semantic information in an iterative manner.

Semantic Segmentation. As an extension of image classification, semantic segmentation assigns pixel-wise labels of object categories for the input image. It is challenging due to randomness of object distribution, color ambiguity,

Table I
COMPARISON BETWEEN VARIOUS INDOOR DATASETS. ITERNET RGB-D IS OUR PROPOSED DATASET. ×: NOT INCLUDED, ✓: INCLUDED, -: RELEVANT INFORMATION NOT AVAILABLE.

Dataset	**NYUv2** [5]	**SUN RGB-D** [6]	**Building Parser** [10]	**Matterport 3D** [11]	**ScanNet** [7]	**SUNCG** [8]	**SceneNet RGB-D** [9]	**IterNet RGB-D**
Year	2012	2015	2017	2017	2017	2017	2016	2019
Type	Real	Real	Real	Real	Real	Synthetic	Synthetic	Synthetic
Images/Scans	1449	10K	70K	194K	1513	130K	5M	12,856
Layouts	464	-	270	90	1513	45,622	57	3214
Object Classes	894	800	13	40	≥ 50	84	255	333
RGB	✓	✓	✓	✓	×	×	✓	✓
Depth	✓	✓	✓	✓	×	✓	✓	✓
Semantic Label	✓	✓	✓	✓	✓	✓	✓	✓
RGB Texturing	Real	Real	Real	Real	Real	Not Photorealistic	Photorealistic	Photorealistic
Image Resolution	640×480	640×480	1080×1080	1280×1024	640×480	640×480	320×240	1280×960; 1280×720

poor illumination, and occlusion. Deng *et al.* [21] proposed a transformation function and a discriminative classifier that maximize the mutual information of data and their labels in the latent space to reduce the uncertainties, *i.e.*, missing and noisy labels. Alterative approaches are typically based on CNNs. Long *et al.* [22] chose to stand on the popular Fully Convolutional Network (FCN) architecture, which inspired most of the subsequent approaches on semantic segmentation. Using the publicly available RGB-D datasets, some methods attempted to use depth information for better segmentation, no longer limited to a single RGB image. Li *et al.* [23] constructed HHA images [24] for the depth channel through geometric encoding before feeding them to the network and used the Long Short-Term Memory (LSTM) network to fuse two different types of features. Ma *et al.* [25] predicted semantic segmentation from RGB-D sequences, but the method is inapplicable to sparse views. Cong *et al.* [26] presented saliency detection with comprehensive information, *e.g.*, depth cue and inter-image correspondence. Our method exploits depth information to help improve semantic segmentation, but the depth is estimated from the input color image instead of directly captured by a dedicated depth sensor. We propose an iterative method for joint estimation of the depth and semantic segmentation, which benefit each other.

Indoor Scene 3D Reconstruction. Indoor Scene 3D Reconstruction from a color video or multi-view color images is a challenging and active topic. Given a color video, most structure from motion (SFM) methods [27] recovered the 3D structure by estimating the motion of the cameras corresponding to the frames. However, it is difficult for these methods to obtain dense and accurate reconstruction. Given multi-view color images with calibrated camera parameters, multi-view stereo (MVS) methods [28] can achieve more accurate 3D reconstruction. But they require adjacent views to have sufficient overlap and cannot work well with sparse views. COLMAP [29], [30] provides a pipeline containing both SFM and MVS with graphical and command-line interfaces. When the views of images are very sparse, the depth of each image can be estimated and fused together using iterative closest point (ICP) like registration methods [31]. However, it is difficult to achieve accurate depth estimation from individual color images which increases the difficulties of ICP fusion. Saxena *et al.* [32] proposed a novel method for 3D reconstruction from sparse views, but it only works well for building-like outdoor scenes and cannot generate semantics. Learning-based methods, *e.g.*, MVSNet [33] and DeepMVS [34], output the depth of a specific frame based on a color multi-view sequence, but they cannot deal with sparse views. In this paper, we design *IterNet* to estimate a more accurate depth map with the help of semantic segmentation, and propose a joint global and local registration method to better achieve indoor scene 3D reconstruction from sparse views.

III. PROPOSED METHOD

In this section, we first introduce our IterNet RGB-D dataset in Section III-A, and then describe the technical details of IterNet for iterative joint depth estimation and semantic segmentation in Section III-B. The joint global and local multi-view reconstruction method is presented in Section III-C. Figure 2 illustrates the workflow of our method.

A. Dataset

Different from other synthetic datasets [8], [9], our dataset is generated by a third-party platform which includes various real-life house styles, real prototype rooms designed by professional designers, and rich material profiles. For fast photorealistic rendering, we use image splitting and recombination to achieve distributed rendering on a cluster of 32 servers. Each server has two CPUs with a total of 32 cores and 64 threads. The average rendering time of a 1280 × 960 image is about 90 seconds. To render 12,856 images, it takes about 321 hours. In terms of rendering quality, in addition to modeling the direct illumination of the light source in the scene, the illumination reflected by other objects, known as Global Illumination (GI), is also taken into consideration. The Brute Force (BF) algorithm [35] based on path tracking is adopted. The number of samples per pixel is up to 512, and varies for different scenes. The noise level (standard deviation) is set below 0.05. A lower noise level would yield better

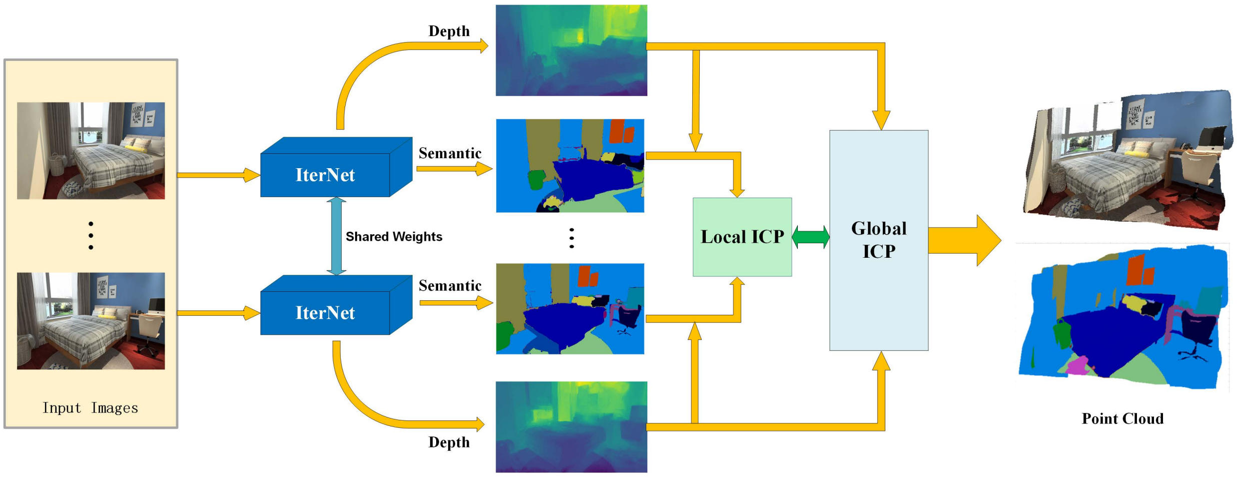


Figure 2. Illustration of the proposed method for indoor 3D reconstruction and understanding. The blue Module refers to our IterNet for iterative joint depth estimation and semantic segmentation (Section III-B). With the help of semantic segmentation, we use our proposed joint global and local registration method to reconstruct a 3D scene with semantic information from sparse views (Section III-C).

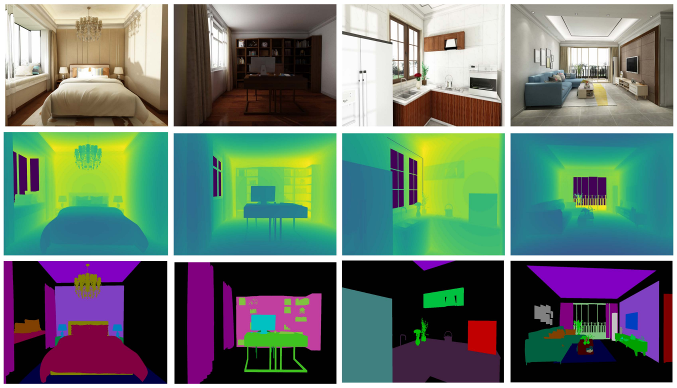


Figure 3. Some examples of different scenarios in our dataset. From top to bottom: color images, ground-truth depth maps, and ground-truth semantic segmentations.

quality, but requires longer rendering time. Rendered images are denoised using a wavelet-based denoising method [36]. Figure 3 shows some examples of different scenarios in our dataset. Our dataset provides photorealistic high-resolution RGB images, accurate depth maps, and pixel-level semantic labels for thousands of layouts. Our dataset will be available online.

B. IterNet: Iterative CNN for Joint Depth Estimation and Semantic Segmentation

Network Architecture. The proposed IterNet is a multi-task deep CNN mainly consisting of two parts: the depth estimation sub-network and the semantic segmentation sub-network, as shown in Figure 4.

Our depth estimation sub-network is inspired by the monocular depth estimation method [20] which uses a continuous conditional random field (CCRF) to combine multi-scale features. Different from [20], we add a semantic branch with an encoder-decoder structure to extract semantic features. Then, multi-scale RGB features and semantic features are integrated by a CCRF module to improve boundary accuracy. The RGB branch consists of a front-end base network and a refinement network combined with several CCRF modules. Together with semantic information, the output of the RGB branch is fed into a CCRF module to generate the depth map which in turn serves as the input of the semantic segmentation sub-network in the next iteration.

In the semantic segmentation sub-network, we use the Long Short-Term Memorized Context Fusion (LSTM-CF) [23] model, which is capable of fusing contextual information from multiple sources (*i.e.* photometric and depth channels). Instead of the original serial vertical and horizontal context layers, we adopt a parallel context layer and a direct fusion scheme to take advantage of the estimated depth map from the RGB branch. An Atrous Spatial Pyramid Pooling (ASPP) [37] is added as a multi-scale feature extractor using multiple parallel filters with different sampling rates. In addition, the depth map is first encoded into an HHA image [24] using geocentric encoding before passing through the segmentation sub-network.

Training and Testing. Given a dataset of *RGB-Depth-Semantic* triplets, our aim is to train the designed network for joint depth and semantic estimation. The depth estimation sub-network and semantic segmentation sub-network are applied alternately to boost the performance. Instead of jointly training the two sub-networks, we train the depth estimation and semantic segmentation sub-networks sequentially for flexible boosting. Taking the depth estimation sub-network as an example, we train the upper branch and the lower branch with *RGB-Depth* pairs and *Sematic-Depth* pairs, respectively. The depth estimation sub-network is then fine-tuned with the *RGB-Depth-Semantic* triplets. The semantic segmentation sub-network is trained in a similar way.

At the test stage, since each sub-network expects the output of the other sub-network as part of input, we use the following strategy. We need an initialized semantic segmentation or depth estimation which can be easily obtained by disabling one of the branches in the original network structure. For example, if we want to obtain an initial depth estimation for semantic segmentation, we disable the semantic segmentation branch

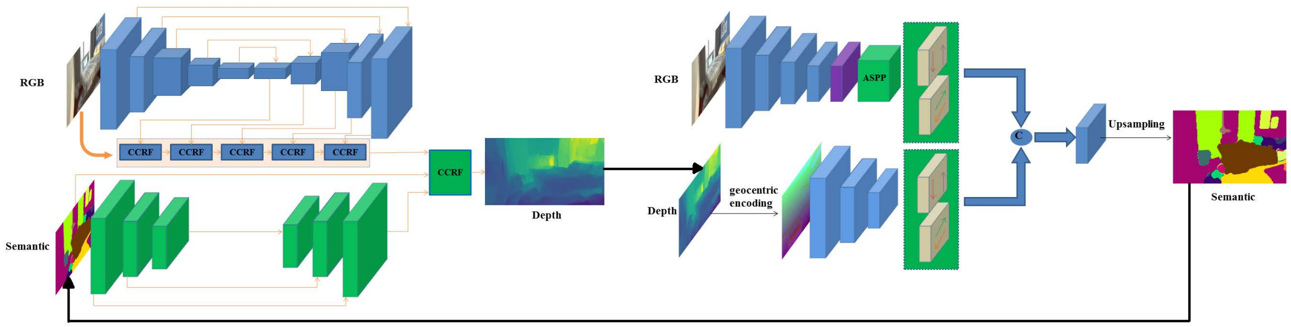


Figure 4. Overview of the proposed IterNet architecture. The CCRF blocks in the depth estimation sub-network fuse features at different scales and combine semantic features. In the semantic segmentation sub-network, the purple block represents the atrous convolution to increase the receptive field. The ASPP (atrous spatial pyramid pooling) block has four different dilated convolutions for resampling in our implementation.

in the depth estimation sub-network and then extract features from RGB branches as an initial depth. We then alternately run the two sub-networks, with the output of one sub-network used as input for the other sub-network. The additional depth information helps improve semantic segmentation, and the semantic segmentation in turn contributes to improved depth estimation. In practice, we find that there is no significant improvement after 3 iterations, which shows quick convergence.

Implementation Details. The proposed approach is implemented on the Caffe framework [38] and runs on a computer with an Nvidia GTX 1080 Ti graphics card (11GB). For the depth estimation sub-network, the learning rate is initialized at 10^{-11} and decreases by 10% for every 30 epochs. The batch size is set to 16. The momentum and the weight decay are set to 0.9 and 0.0005, respectively. The semantic segmentation sub-network follows the same training rules, but the learning rate is initialized at 10^{-4}. The parameters of batch size, momentum and weight decay are set to 8, 0.9 and 0.005, respectively. The learning rate decreases by 10% for every 20 epochs. For the fine-tuning of each sub-network, the initial learning rates are set to 10^{-12} and 10^{-5} for depth estimation and semantic segmentation, respectively. The batch size, momentum and weight decay remain the same as the pretraining.

C. Joint Global and Local Reconstruction

After depth estimation and semantic segmentation for the image of each view, we reconstruct the whole 3D scene by fusing the depth maps of different views. The straightforward way is to use the ICP algorithm to align the partial 3D models from different perspectives, but would be difficult to achieve satisfactory alignment. First, the depth maps are obtained by a monocular depth estimation network instead of depth sensors like Kinect, and contain significant amount of non-statistical errors. It is therefore insufficient to align two depth point clouds with just one rigid transformation. Second, for sparse perspectives, two adjacent views have limited overlap, and are difficult to align by the standard ICP algorithm. Hence, we propose a new joint global and local registration method by exploiting photometric and semantic information to improve the reconstruction quality.

Before 3D fusion, we filter the noisy 3D points based on the plane constraint similar to [39]. After depth estimation and semantic segmentation, each view now contains three components: color image C_i, depth map D_i, and semantic segmentation S_i, where i denotes view index. Let $\mathcal{X} \triangleq \{X_i\} = \{(C_i, D_i, S_i)\}_{i=1}^N$ be the set of color-depth-semantic triplets for N sparse views. We align all the depth point clouds sequentially with the previous registration result used as the next target model. Each alignment has two stages, namely the global alignment and local alignment.

Global Alignment. Taking the point cloud reconstructed from the previous $i-1$ views as the target, our goal for global alignment is to find an optimal global rigid transformation $\mathcal{T}_i = \{R_i, t_i\}$ for view i, where R_i and t_i denote the rotation and translation, respectively. Specifically, we first convert the depth map D_i into a point cloud denoted by $\mathcal{P}_i = \{p_j\}_{j=1}^{n_i}$, where p_j denotes a 3D point and n_i represents the total number of points at view i. Our global registration stands on the ICP-type framework which alternates two steps until convergence. The transformation is initialized with a 4×4 identity matrix. Let $\mathcal{P}'_{i-1} = \{p'_j\}_{j=1}^{n'_{i-1}}$ be the target point cloud containing all the n'_{i-1} fused points from previous $i-1$ views, the first step finds for each point $p_k \in \mathcal{P}_i$ its corresponding point $p'_j \in \mathcal{P}'_{i-1}$ if possible, and the second step optimizes the transformation $\mathcal{T}_i$ by aligning the point cloud $\mathcal{P}_i$ at view i to the fused point cloud $\mathcal{P}'_{i-1}$.

In the first step, we incorporate the additional photometric and semantic information for more accurate matching. To this end, we lift each point $p_k \in \mathcal{P}_i$ from the 3D space to a 7-dimensional (7D) point $\hat{p}_k \in \hat{\mathcal{P}}_i$ in a 7D space by including color and semantic information. Specifically, $\hat{p}'_k \triangleq \{\hat{p}'_k(s)\}_{s=1}^7$ include the 3D position $\{\hat{p}_k(s)\}_{s=1}^3$, RGB color $\{\hat{p}_k(s)\}_{s=4}^6$, and semantic label $\hat{p}_k(7)$. Similarly, the point $p'_j \in \mathcal{P}'_{i-1}$ is lifted to a 7D point $\hat{p}'_j \in \hat{\mathcal{P}}'_{i-1}$. Our global registration method first finds the corresponding point $\hat{p}'_{\tilde{k}} \in \hat{\mathcal{P}}'_{i-1}$ for each point $\hat{p}_k$ in $\hat{\mathcal{P}}_i$ by the following optimization:

$$\tilde{k} = \mathop{\arg\min}_{j \in \{1,2,\ldots,n'_{i-1}\}} \|\hat{p}_k - \hat{p}'_j\|_w^2. \tag{1}$$

The weighted Euclidean distance, $\|\hat{p}_k - \hat{p}'_j\|_w$, between $\hat{p}_k$ and

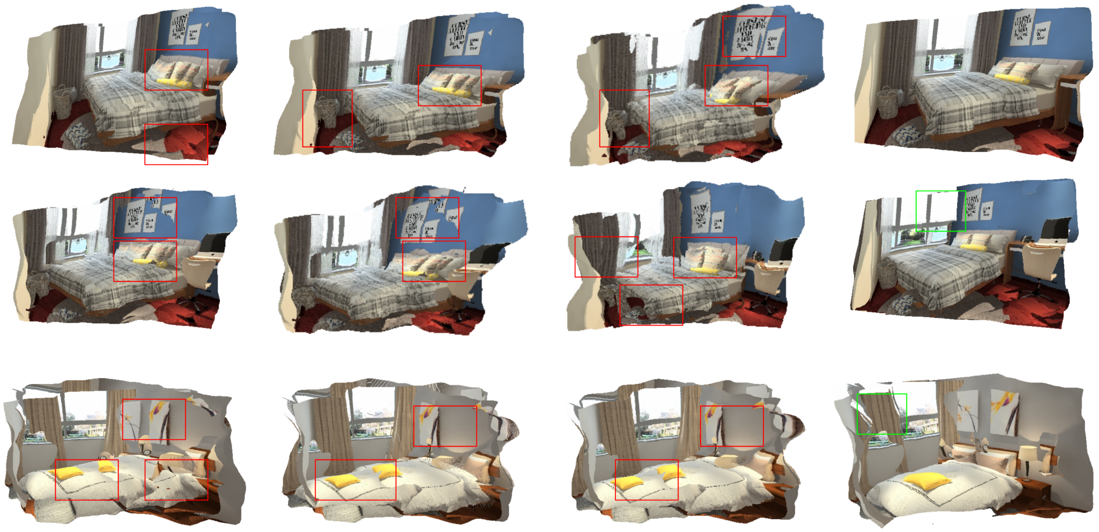


Figure 5. Comparison of different alignment methods: From left to right are results of standard ICP algorithm [31], 4PCS [40], global alignment using the estimated depth without the help of semantic branch, and our joint global and local alignment method.

$\hat{p}'_j$ in the 7D space is defined as

$$\begin{aligned}\|\hat{p}_k - \hat{p}'_j\|_w^2 = &\sum_{s=1}^{3}(\hat{p}_k(s) - \hat{p}'_j(s))^2 \\ &+ w_1\sum_{s=4}^{6}(\hat{p}_k(s) - \hat{p}'_j(s))^2 \\ &+ w_2(\hat{p}_k(7) - \hat{p}'_j(7))^2,\end{aligned} \tag{2}$$

where w_1 and w_2 are weights to balance the importance of geometric, photometric and semantic information. They are set to be $w_1 = 0.1$ and $w_2 = 10$ in our experiments.

Due to limited overlap, not all the points in $\hat{\mathcal{P}}_i$ have their corresponding points in $\hat{\mathcal{P}}'_{i-1}$. We reject $\hat{p}'_{\tilde{k}}$ if the matching error is larger than a threshold. In our implementation, this threshold is set to 5cm. Let $\mathcal{C}_i = \{p_k, \hat{p}'_{\tilde{k}}\}$ be the set of retained correspondences. In the second step, since photometric and semantic matching errors are independent of rigid transformations, we use a standard ICP algorithm [31] to find the transformation between the two point clouds:

$$(R_i, t_i) = \arg\min_{R,t} \sum_{(p_k, \hat{p}'_{\tilde{k}}) \in \mathcal{C}_i} \|\hat{p}'_{\tilde{k}} - R\hat{p}_k - t\|_2^2. \tag{3}$$

Local Alignment. The global registration on the 7D feature space is able to achieve coarse alignment, but still cannot handle non-rigid local deformation, *e.g.*, due to non-statistical errors in monocular depth estimation. To address this problem, we further propose a local registration strategy to refine the previous coarse estimation, similar to coarse-to-fine refinement. Specifically, we first extract local point sets from the original point cloud according to their semantic labels, and then register each of them using the above method. Note that in this case, a subset of points from one view is only matched to subsets of points with the same semantic label. Therefore, when finding the matched point, the semantic difference term in Eq. (2) is always zero. For each local set, once it is aligned, we fuse the registered parts from different views by averaging 3D positions of overlaps to mitigate the influence of noise. The key for our joint global and local registration method is to use multiple transformations to register sparse views with coarse-to-fine refinement, rather than just one single transformation, which is more robust to the noise and outliers in the monocular depth estimation.

IV. Experimental Results

A. Ablation Study

We compare the full model with the one without semantic segmentation and the one without depth estimation in Table II. It can be seen that the full model has achieved the best performance. Figure 5 shows the fusion results of an ICP matching method [31], 4PCS [40], the global alignment using the estimated depth without the help of semantic branch, and our proposed joint global and local registration method. The red boxes highlight the areas that are difficult to align with other methods, and the green boxes indicate features lost by our method. The results show that some misalignments occur in local areas for standard ICP methods. Our joint global and local alignment method cannot ensure that all the features are captured and aligned, such as the curtain on both sides of the window (in the green boxes). However, compared with other alignment methods, our method achieves better fusion result in terms of both global structure and local details.

Our iterative scheme in IterNet usually converges to promising results after three iterations and is stable for various images. Figure 6 shows the rapid decreasing of average RMS (root mean squared) errors of depth estimation over all the test images with respect to the number of iterations. The average pixel accuracy of semantic segmentation over all the test images also increases sharply to the stationary point in three iterations. No significant improvement for both depth estimation and semantic segmentation is observed beyond three iterations.

To study and verify the effectiveness of IterNet in depth estimation, we compare it with two recent backbone archi-

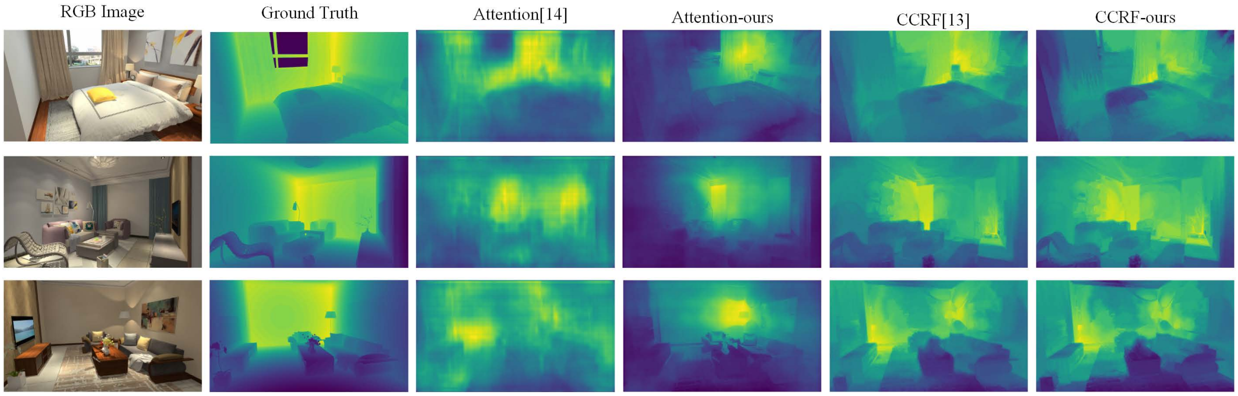


Figure 7. Comparison of depth estimation with two different network architectures.

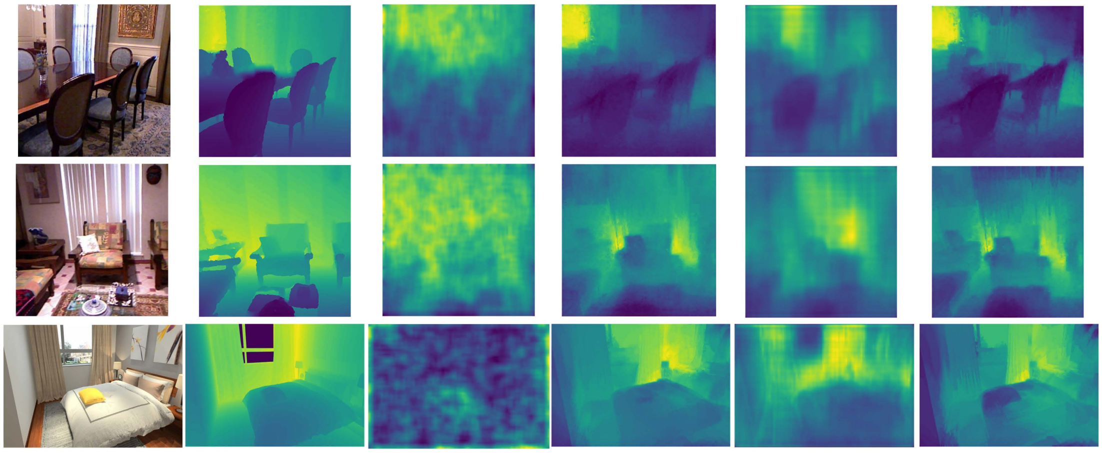


Figure 8. Depth estimation results on NYUv2 dataset (top two rows) and our dataset (bottom row). From left to right are the input RGB images, the ground-truths depth and the depth results estimated by Eigen *et al.* [16], Xu *et al.* [20], Xu and Wang [19], and our method.

Table II
ABLATION RESULTS ON OUR DATASET. F-S: FULL MODEL WITHOUT SEMANTIC; F-D: FULL MODEL WITHOUT DEPTH; F: FULL MODEL.

Method	F-S	F-D	F
rel (lower is better)	0.176	-	**0.136**
log10 (lower is better)	0.088	-	**0.062**
rms (lower is better)	1.012	-	**0.507**
P-acc.(%) (higher is better)	-	67.35	**75.54**
M-acc.(%) (higher is better)	-	68.29	**74.49**
IoU(%) (higher is better)	-	54.21	**63.98**

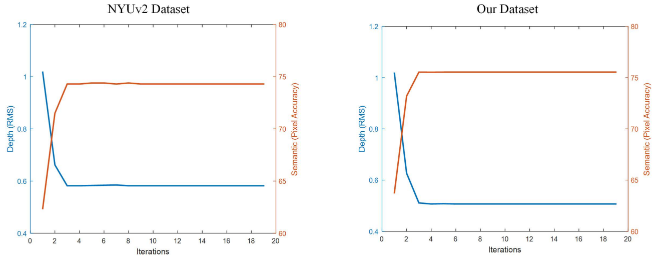


Figure 6. Convergence curves of the proposed IterNet for NYUv2 dataset [5] and our dataset (averaged over all test images in each dataset).

tectures including Structured Attention Guided Convolution Neural Fields [19] and CCRF [20] which achieve promising performance in depth estimation. Figure 7 shows the comparison results on our IterNet RGB-D dataset. High resolution images are cropped into small pieces of 426 × 426, and are fed into the networks. The results show that our framework significantly enhances the attention-based network with clear object structures, and refines the CCRF architecture with sharper contours for some objects such as the pillow and the chair.

B. Depth Estimation

We compare our approach with several state-of-the-art methods on NYUv2 dataset [5] in Table IV. We use 795 images for training and other 654 images for testing as other methods did. We also use the same raw data as other methods and adopt data augmentation (finally 4770 images for training) to avoid the over-fitting problem. Let d_i and d_i^* denote the predicted depth value and the ground-truth value for pixel i, respectively, and P represent the total number of pixels. Referring to previous work [16], [17], [41], we evaluate the depth estimation results with the following metrics: (1) mean relative error (rel): $\frac{1}{P}\sum_i \frac{|d_i - d_i^*|}{d_i^*}$; (2) root mean squared error (rms): $\sqrt{\frac{1}{P}\sum_i (d_i - d_i^*)^2}$; (3) mean log10 error (log10):

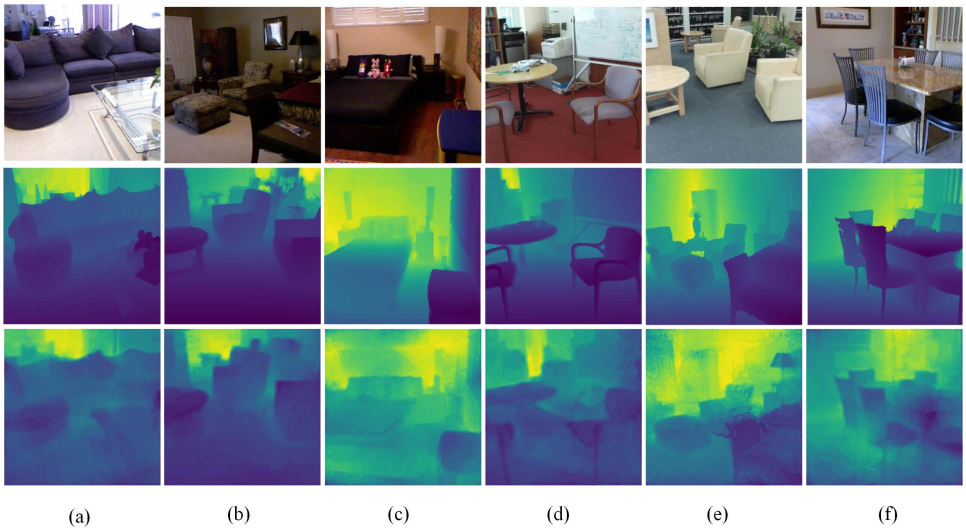


Figure 9. Depth estimation results on NYUv2 dataset [5] (a, b, c) and SUN RGB-D dataset [6] (d, e, f) using our model trained by our dataset. From top to bottom are the input color images, the ground truths, and our estimated depths.

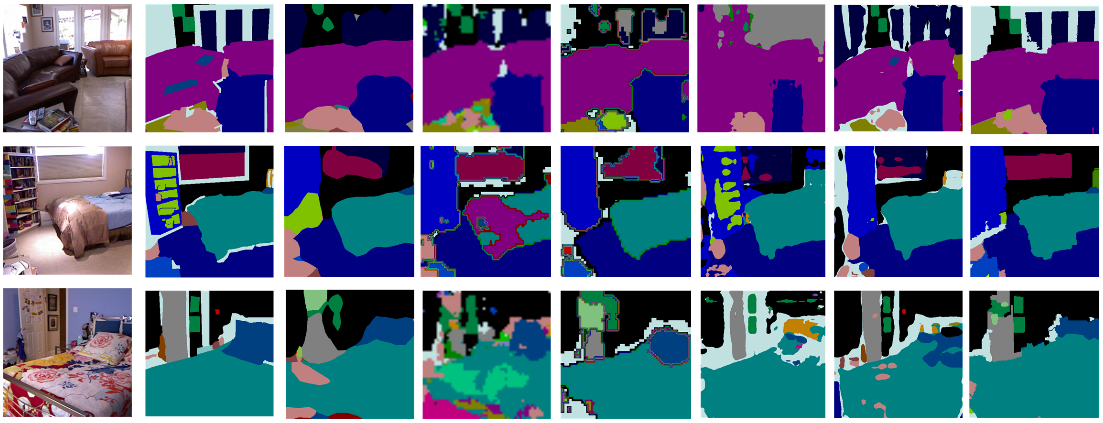


Figure 10. Semantic segmentation results on NYUv2 dataset (top two rows) and our dataset (bottom row). From left to right are the input RGB images, the ground-truths and the results estimated by FCN [22], Chen *et al.* [37], Li *et al.* [23], Zhao *et al.* [50], DANet [51] and our method.

$\frac{1}{P}\sum_i \|\log_{10}(d_i)-\log_{10}(d_i^*)\|$; and (4) accuracy with threshold t: percentage (%) of d_i^* subject to $\delta \triangleq \max(\frac{d_i^*}{d_i}, \frac{d_i}{d_i^*}) < t$. Our results are averaged over three test trials. The results of the compared methods are quoted from their papers. As shown in Table IV, our method outperforms thirteen competing methods in all metrics, and is comparable to PAD-Net [18] which has a more complex network structure and requires ground-truth contours and normals as part of labels.

We also quantitatively evaluate some methods with their provided code on our IterNet RGB-D dataset. As shown in Table VI, our method achieves the most accurate depth estimation on all the metrics. Figure 8 gives some visual comparison results on NYUv2 dataset [5] and our dataset.

Figure 9 shows some depth estimation results for real indoor scenes from the NYUv2 dataset [5] and SUN RGB-D dataset [6] without finetuning. The results demonstrate that our model trained on our dataset has promising generalizability to other datasets.

Table III reports the running time results of our proposed depth estimation module and other competing methods. All the methods are run on the same desktop equipped with an Intel Xeon 2.10GHz CPU, 64GB RAM, and an NVIDIA GeForce GTX 1080Ti GPU. Requiring three iterations for convergence, our network takes longer time particularly at the GPU mode.

Table III
RUNNING TIMES OF DIFFERENT DEPTH ESTIMATION METHODS.

Mode	Eigen *et al.* [16]	Laina *et al.* [42]	Xu *et al.* [20]	Xu and Wang [19]	Ours
CPU	10.63s	11.36s	21.56s	32.20s	35.53s
GPU	11.99s	7.27s	6.01s	5.02s	21.47s

C. Semantic Segmentation

To evaluate the performance of semantic segmentation, we use NYUv2-40 dataset [22] in which all objects in the NYUv2 dataset [5] are divided into 40 categories. We use the same training and testing data as other methods and adopt three metrics in percentage (%): pixel accuracy, mean accuracy, and Intersection over Union (IoU). As shown in Table VII, our inferred semantic segmentation results outperform those state-of-the-art methods. We also quantitatively evaluate some recent work that provide source code on our IterNet RGB-D dataset. Results in Table VIII show that our method also achieves the best performance. Figure 10 presents some visual

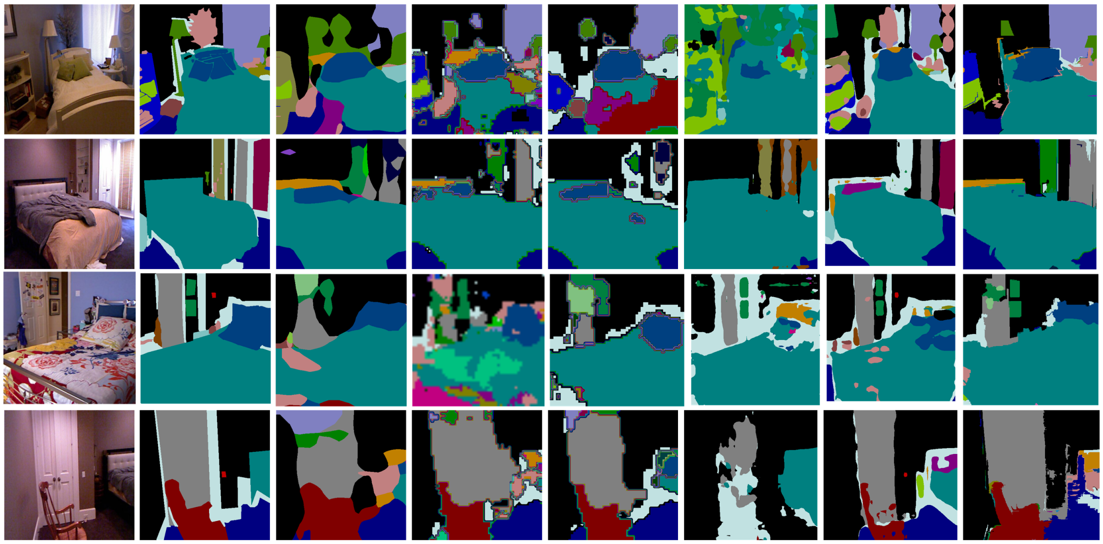


Figure 11. Semantic segmentation results on NYUv2 dataset [5]. From left to right are the input RGB images, the ground-truths and the results estimated by FCN [22], Chen *et al.* [37], Li *et al.* [23], Zhao *et al.* [50], DANet [51] and our method.

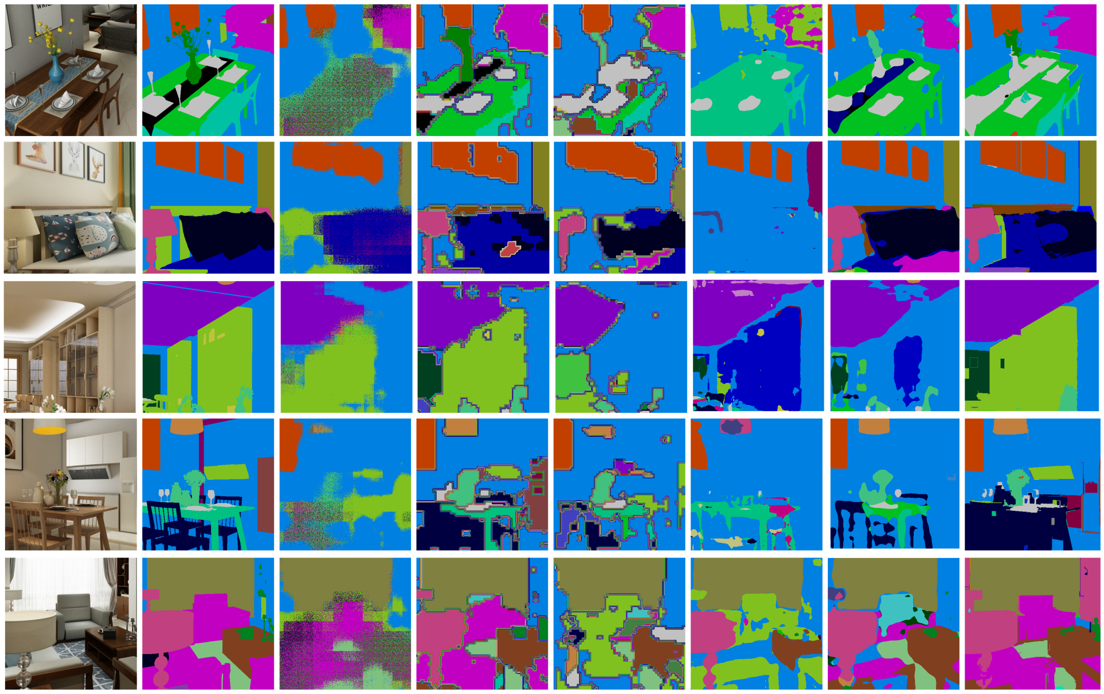


Figure 12. Semantic segmentation results on our dataset. From left to right are the input RGB images, the ground-truths and the results estimated by FCN [22], Chen *et al.* [37], Li *et al.* [23], Zhao *et al.* [50], DANet [51] and our method.

comparison results on NYUv2-40 dataset and our dataset mapped into 87 categories. Being consistent with the quantitative results in Table VII and Table VIII, our approach generates more accurate semantic segmentation results on both real dataset (NYUv2) and synthetic dataset (IterNet RGB-D) than state-of-the-art methods. More visual comparison results for semantic segmentation in Figure 11 and Figure 12 show that our approach generates more accurate semantic segmentation on both real dataset (NYUv2) and synthetic dataset (IterNet RGB-D) than other four competing methods.

Table V reports the running times of different semantic segmentation methods. Similar to the depth estimation module, the semantic segmentation module also takes longer time, but still at the same order of magnitude. As future work, we will investigate more efficient network architectures (e.g. lightweight network modules) to reduce the running times.

D. Multi-view Reconstruction

In Figure 13, we evaluate multi-view 3D reconstruction performance of the proposed method on NYUv2 dataset [5] and our dataset using three wide-baseline views, compared with four state-of-the-art multi-view stereo methods: COLMAP [29], [30], PMVS2 [57], OpenMVS [58], and DeepMVS [34].

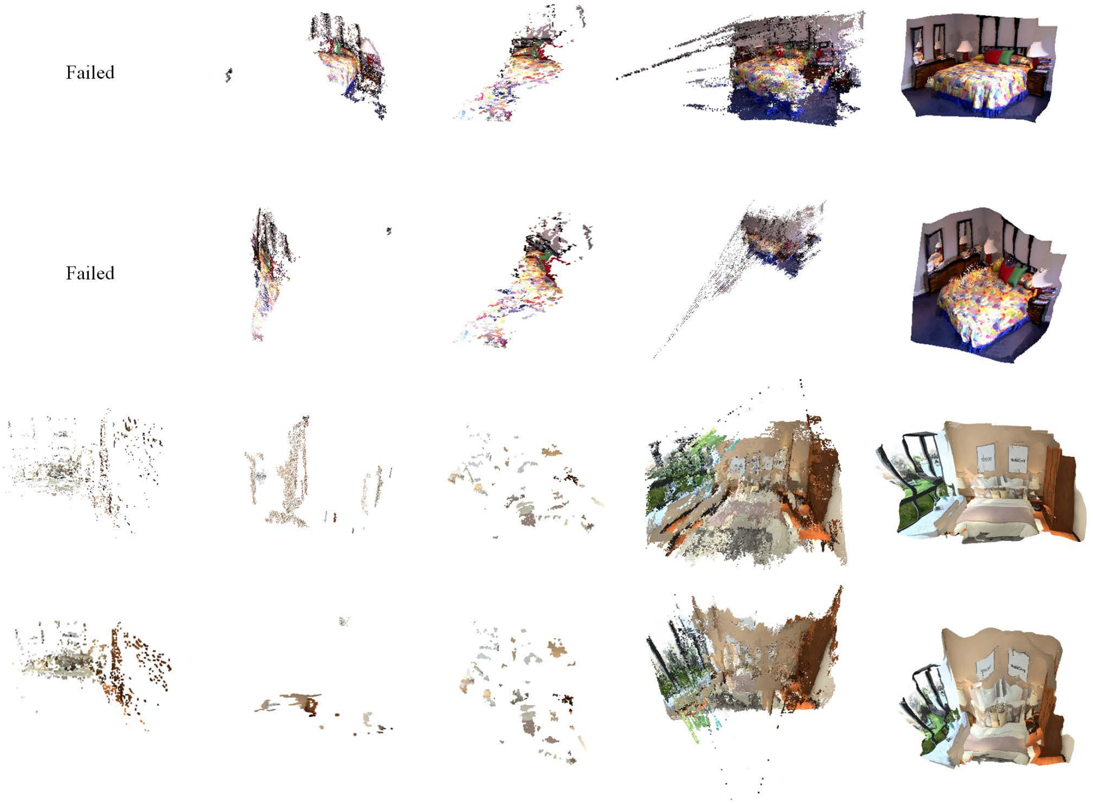


Figure 13. Comparison of scene reconstruction results of different methods on NYUv2 dataset (top two rows) and our dataset (bottom two rows). From left to right are the results of COLMAP [29], [30], PMVS2 [57], OpenMVS [58], DeepMVS [34] and our method.

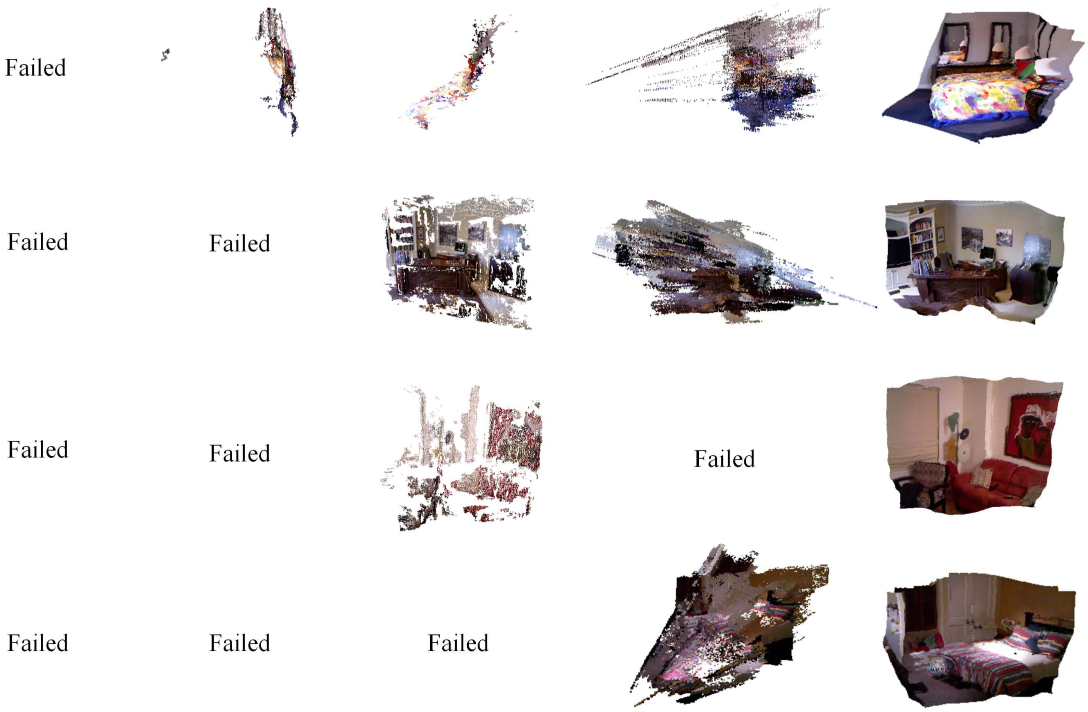


Figure 14. Comparison of multi-view reconstruction results of different methods on NYUv2 dataset [5]. From left to right are the results of COLMAP [29], [30], PMVS2 [57], OpenMVS [58], DeepMVS [34] and our method.

Table IV
QUANTITATIVE EVALUATION FOR DEPTH ESTIMATION ON NYUV2 DATASET.

Method	Error (lower is better)			Accuracy (higher is better)		
	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Saxena *et al.* [43]	0.349	-	1.214	0.447	0.745	0.897
Liu *et al.* [15]	0.335	0.127	1.06	-	-	-
Karsch *et al.* [44]	0.35	0.131	1.20	-	-	-
Ladicky *et al.* [45]	-	-	-	0.542	0.829	0.941
Zhou *et al.* [46]	0.305	0.122	1.04	0.525	0.838	0.962
Liu *et al.* [47]	0.213	0.087	0.759	0.650	0.906	0.976
Roi and Todorovic [48]	0.187	0.078	0.744	-	-	-
Eigen *et al.* [16]	0.215	-	0.907	0.611	0.887	0.971
Eigen and Fergus [41]	0.158	-	0.641	0.769	0.950	**0.988**
Laina *et al.* [42]	0.129	0.056	0.583	0.801	0.950	0.986
Xu *et al.* [20]	0.139	0.063	0.609	0.793	0.948	0.984
Xu and Wang [19]	0.121	0.052	0.586	0.811	**0.954**	0.987
Joint HCRF [17]	0.220	0.094	0.745	0.605	0.890	0.970
Jafari *et al.* [49]	0.157	0.068	0.673	0.762	0.948	**0.988**
PAD-Net [18]	**0.120**	0.055	**0.582**	0.817	**0.954**	0.987
Ours	0.122	**0.051**	**0.582**	**0.819**	0.953	**0.988**

Table V
RUNNING TIMES OF DIFFERENT SEMANTIC SEGMENTATION METHODS.

Mode	FCN [22]	Chen *et al.* [37]	Li *et al.* [23]	Zhao *et al.* [50]	Ours
CPU	5.85s	7.71s	7.79s	10.56s	35.53s
GPU	2.65s	1.33s	4.69s	7.47s	21.47s

We obtain the sparse views for NYUv2 dataset by selecting one frame per 30-40 frames, and use the camera parameters estimated by COLMAP [29] for OpenMVS [58], PMVS2 [57] and DeepMVS [34]. As shown in Figure 13, COLMAP [29], [30] fails to generate meaningful results on NYUv2 dataset from sparse views. Incorrect 3D points in results reconstructed by PMVS2 [57] and OpenMVS [58] are observed: some points gather together from side view and top view on NYUv2 dataset. Moreover, their obtained point clouds are too sparse to provide acceptable results even enhanced by linear interpolation. DeepMVS reconstructs more points compared with the traditional methods, but the reconstructed model contains a lot of noise and outliers. On the contrary, our method achieves the best results for sparse multi-view reconstruction by considering 7D information (geometry, photometry and semantics) and by using joint global and local registration. More results on NYUv2 dataset [5] and our dataset using three or four sparse views are given in Figure 14 and Figure 15, respectively. We observe that the multi-view stereo method in COLMAP [30] fails to generate 3D point clouds, and the point clouds reconstructed by OpenMVS [58] and PMVS2 [57] lack sufficient density and completeness. Although DeepMVS [34] achieves dense reconstruction, the reconstructed model contains many incorrect points. In contrast, our method achieves accurate and complete reconstruction from sparse views. Because COLMAP [30] fails for most scenes in NYUv2 dataset [5], we give quantitative evaluation on our dataset in Table IX. Two metrics are used to evaluate the results of MVS reconstruction: accuracy and completeness. Accuracy represents the average distance between the points on reconstructed model and the nearest points on the ground-truth model. Completeness measures the percentage of the points on the ground-truth model that can find corresponding points on the reconstructed model within a certain distance threshold (0.1). We generate the 3D ground-truth model by fusing multi-view ground-truth depth point clouds using ICP. As shown in Table IX, our method achieves the most complete 3D reconstruction, significantly outperforming other competing methods. Although traditional multi-view stereo methods [30], [57], [58] have higher accuracy, their reconstructed points are too sparse to provide acceptable results even enhanced by linear interpolation. Figure 16 shows our reconstructed models on NYUv2 dataset [5] and our dataset presented from five different views.

Table VI
QUANTITATIVE EVALUATION FOR DEPTH ESTIMATION ON OUR DATASET.

Method	Error (lower is better)			Accuracy (higher is better)		
	rel	log10	rms	$\delta < 1.15$	$\delta < 1.15^2$	$\delta < 1.15^3$
Eigen *et al.* [16]	0.948	0.285	4.711	0.054	0.205	0.492
Laina *et al.* [42]	0.404	0.235	3.433	0.102	0.310	0.581
Xu *et al.* [20]	0.175	0.089	1.010	0.435	0.700	0.907
Xu and Wang [19]	0.151	0.067	0.620	0.536	0.817	0.975
Ours	**0.136**	**0.062**	**0.507**	**0.568**	**0.918**	**0.982**

Table VII
QUANTITATIVE EVALUATION FOR SEMANTIC SEGMENTATION ON THE NYUV2-40 DATASET.

Method	Pixel Accuracy	Mean Accuracy	IoU
Deng *et al.* [52]	63.8	31.5	-
FCN [22]	60.0	42.2	29.2
FCN-HHA [22]	65.4	46.1	34.0
Eigen *et al.* [41]	65.6	45.1	34.1
Lin *et al.* [53]	70.0	53.6	40.6
RefineNet [54]	73.6	58.9	46.5
Kong *et al.* [55]	72.1	-	44.5
Saxena *et al.* [43]	-	55.7	43.1
Gupta *et al.* [24]	60.3	-	28.6
Mousavian *et al.* [56]	68.6	52.3	39.2
DANet [51]	73.9	59.1	47.9
Ours	**74.3**	**59.4**	**48.7**

Table VIII
QUANTITATIVE EVALUATION FOR SEMANTIC SEGMENTATION ON OUR DATASET.

Method	Pixel Accuracy	Mean Accuracy	IoU
FCN [22]	47.07	33.76	24.63
Chen *et al.* [37]	66.28	67.98	53.90
Li *et al.* [23]	61.97	46.93	40.46
Zhao *et al.* [50]	74.82	72.36	60.91
DANet [51]	75.35	74.22	63.87
Ours	**75.54**	**74.49**	**63.98**

V. CONCLUSIONS

In this paper, we solve a challenging problem: reconstructing and understanding indoor 3D scenes based on several

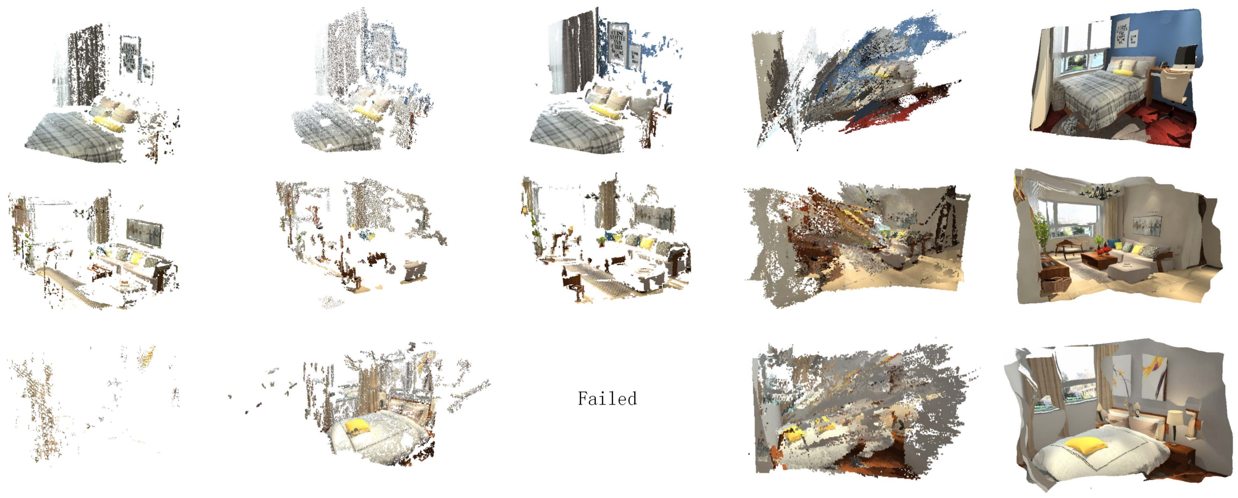


Figure 15. Comparison of scene reconstruction results of different methods on our dataset. From left to right are the results of COLMAP [29], [30], PMVS2 [57], OpenMVS [58], DeepMVS [34] and our method.

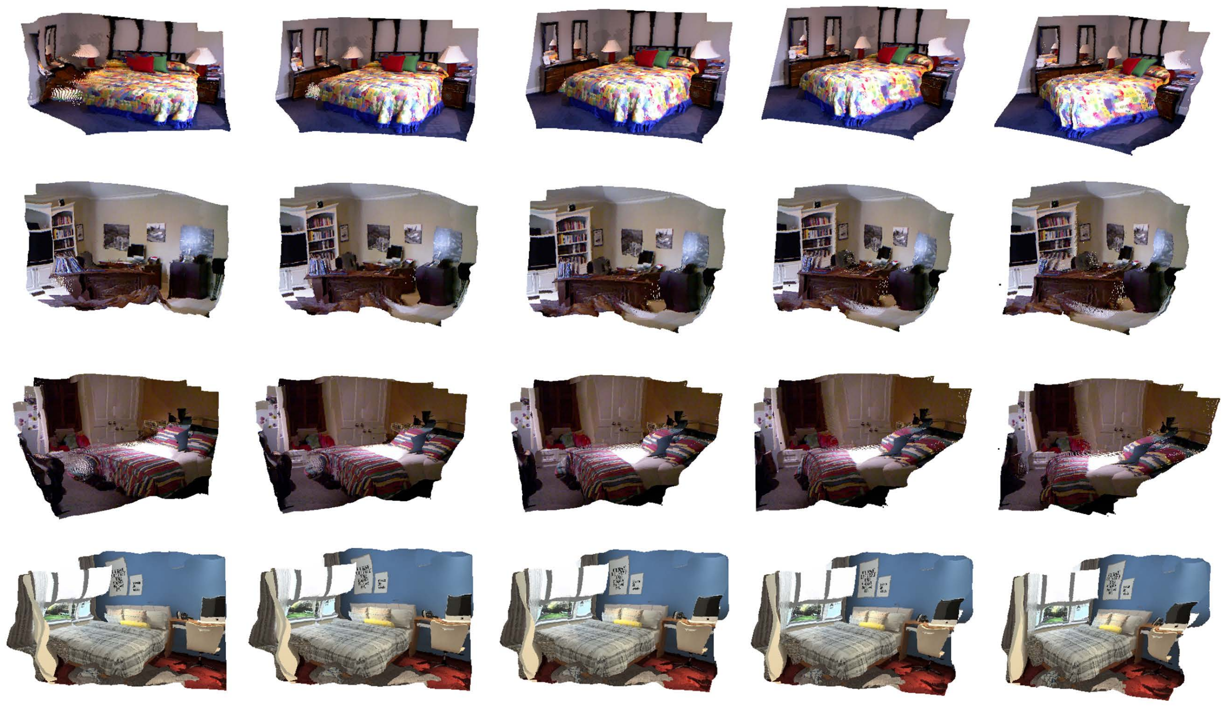


Figure 16. Our reconstructed models on NYUv2 dataset [5] and our dataset presented from five different views as illustrated for each scene.

Table IX
QUANTITATIVE EVALUATION FOR MULTI-VIEW RECONSTRUCTION.

Method	Accuracy (lower is better)	Completeness (higher is better)
COLMAP [30]	3.74	2.33%
PMVS2 [57]	3.71	1.83%
OpenMVS [58]	3.68	1.25%
DeepMVS [34]	21.49	12.47%
Ours	17.72	31.55%

color images captured from uncalibrated sparse views. A novel iterative network, IterNet, is proposed to jointly estimate depth map and semantic segmentation from a single color image. We proposed a joint global and local registration method to reconstruct indoor 3D scenes from sparse views. We also build and make available the IterNet RGB-D dataset, a new dataset that simultaneously provides high-resolution photorealistic RGB images, accurate depth maps, and pixel-level semantic labels for thousand of layouts. Experimental results on both public datasets and our dataset demonstrate that our method achieves the best results on depth estimation, semantic segmentation and multi-view reconstruction, compared with state-of-the-art methods.

References

[1] 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 *ISMAR*, pages 127–136, 2011.

[2] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Trans. Graphics*, 36(4):76a, 2017.

[3] Zhibin Liu, Zongying Shi, and Wenli Xu. On optimal dynamic sequential search for matching in real-time machine vision. *IEEE Trans. Image Processing*, 19(11):3000–3011, 2010.

[4] Hainan Cui, Shuhan Shen, Wei Gao, and Zhanyi Hu. Efficient large-scale structure from motion by fusing auxiliary imaging information. *IEEE Trans. Image Processing*, 24(11):3561–3573, 2015.

[5] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

[6] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas A Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, volume 2, page 10, 2017.

[8] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *CVPR*, 2017.

[9] Ankur Handa, Viorica Pătrăucean, Simon Stent, and Roberto Cipolla. Scenenet: An annotated model generator for indoor scene understanding. In *ICRA*, pages 5737–5743, 2016.

[10] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[11] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[12] Muzammal Naseer, Salman H Khan, and Fatih Porikli. Indoor scene understanding in 2.5/3D: A survey. *arXiv preprint arXiv:1803.03352*, 2018.

[13] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. *ACM Trans. Graphics*, 24(3):577–584, 2005.

[14] Weicheng Huang, Xun Cao, Ke Lu, Qionghai Dai, and Alan Conrad Bovik. Toward naturalistic 2D-to-3D conversion. *IEEE Trans. Image Processing*, 24(2):724–733, 2015.

[15] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *CVPR*, pages 716–723, 2014.

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014.

[17] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, pages 2800–2809, 2015.

[18] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *arXiv preprint arXiv:1805.04409*, 2018.

[19] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, pages 3917–3925, 2018.

[20] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017.

[21] Yue Deng, Feng Bao, Xuesong Deng, Ruiping Wang, Youyong Kong, and Qionghai Dai. Deep and structured robust information theoretic learning for image analysis. *IEEE Trans. Image Processing*, 25(9):4209–4221, 2016.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[23] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. LSTM-CF: Unifying context modeling and fusion with lstms for RGB-D scene labeling. In *ECCV*, pages 541–557, 2016.

[24] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, pages 345–360, 2014.

[25] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with RGB-D cameras. In *IROS*, 2017.

[26] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming Ming Cheng, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits & Systems for Video Technology*, PP(99):1–1, 2018.

[27] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics*, 25(3):835–846, 2006.

[28] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Trans. VCG*, 16(3):407–418, 2010.

[29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016.

[30] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016.

[31] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[32] A Saxena, Sun Min, and A. Y Ng. 3-d reconstruction from sparse views using monocular vision. In *ICCV*, 2007.

[33] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018.

[34] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *CVPR*, pages 2821–2830, 2018.

[35] Ray Tracing. Distributed ray tracing. *ACM Siggraph Computer Graphics*, 18(3):137–145, 1984.

[36] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[37] 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 Trans. PAMI*, 40(4):834–848, 2018.

[38] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678, 2014.

[39] András Bódis-Szomorú, Hayko Riemenschneider, and Luc Van Gool. Superpixel meshes for fast edge-preserving surface reconstruction. In *CVPR*, pages 2011–2020, 2015.

[40] D. Aiger, N. J. Mitra, and D. Cohen-Or. 4-points congruent sets for robust surface registration. *ACM Trans. Graphics*, 27(3):#85, 1–10, 2008.

[41] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015.

[42] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248, 2016.

[43] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Trans. PAMI*, 31(5):824–840, 2009.

[44] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans. PAMI*, 36(11):2144–2158, 2014.

[45] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, pages 89–96, 2014.

[46] Wei Zhuo, Mathieu Salzmann, Xuming He, and Miaomiao Liu. Indoor scene structure analysis for single image depth estimation. In *CVPR*, pages 614–622, 2015.

[47] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. PAMI*, 38(10):2024–2039, 2016.

[48] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, pages 5506–5514, 2016.

[49] Omid Hosseini Jafari, Oliver Groth, Alexander Kirillov, Michael Ying Yang, and Carsten Rother. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation. In *ICRA*, pages 4620–4627, 2017.

[50] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018.

[51] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2018.

[52] Zhuo Deng, Sinisa Todorovic, and Longin Jan Latecki. Semantic segmentation of RGBD images with mutex constraints. In *ICCV*, pages 1733–1741, 2015.

[53] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, pages 3194–3203, 2016.

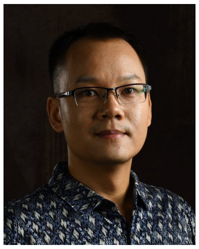
[54] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[55] Shu Kong and Charless Fowlkes. Recurrent scene parsing with perspective understanding in the loop. *arXiv preprint arXiv:1705.07238*, 2017.

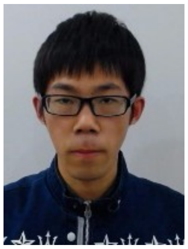
[56] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *3DV*, pages 611–619, 2016.

[57] Y Furukawa and J Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. PAMI*, 32(8):1362–1376, 2010.

[58] OpenMVS. Open multi-view stereo reconstruction library. http://cdcseacave.github.io/openMVS.



Jingyu Yang (M'10-SM'17) received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and Ph.D. (Hons.) degree from Tsinghua University, Beijing, in 2009. He has been a Faculty Member with Tianjin University, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. He was with Microsoft Research Asia (MSRA) in 2011, and the Signal Processing Laboratory, EPFL, Lausanne, Switzerland, in 2012, and from 2014 to 2015. His research interests include image/video processing, 3D imaging, and computer vision.



Ji Xu received the B.E. degree from the School of Information Engineering, Yangzhou University, Yangzhou, China, in 2017. He is currently pursuing the M.E. degree at the College of Electrical and Information Engineering, Tianjin University, Tianjin, China. His interests are mainly in 3D reconstruction and computer vision.

Kun Li received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She visited École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. She is currently an Associate Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. Her research interests include dynamic scene 3D reconstruction and image/video processing.

Yu-Kun Lai received his bachelor and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a Reader of Visual Computing in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of Computer Graphics Forum and The Visual Computer.

Huanjing Yue (M'17) received the B.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2010 and 2015, respectively. She was an Intern with Microsoft Research Asia from 2011 to 2012, and from 2013 to 2015. She visited the Video Processing Laboratory, University of California at San Diego, from 2016 to 2017. She is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. Her current research interests include image processing and computer vision.

Jianzhi Lu received the B.E degree from the College of Informatics, South China Agricultural University in 2012. He is currently serving for the Cloud rendering lab of 3VJ in Guangzhou, China. His research interests include distributed computing, system integration and rendering.

Hao Wu received the B.S. degree in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2011. He is currently working toward the Ph.D. degree in electrical engineering in Tianjin University, Tianjin, China. His research interests include image/video compression and computer vision.

Yebin Liu received the BE degree from Beijing University of Posts and Telecommunications, China, in 2002, and the PhD degree from the Department of Automation, Tsinghua University, Beijing, China, in 2009. He has been working as a research fellow at the computer graphics group of the Max Planck Institute for Informatik, Germany, in 2010. He is currently an associate professor in the Department of Automation, Tsinghua University. His research areas include computer vision, computer graphics and computational photography.