Towards High-Fidelity Single-view Holistic Reconstruction of Indoor Scenes

Haolin Liu[1,2⋆], Yujian Zheng[1,2⋆], Guanying Chen[1,2], Shuguang Cui[1,2], and Xiaoguang Han[1,2†]

[1] School of Science and Engineering, CUHK-Shenzhen
[2] The Future Network of Intelligence Institute, CUHK-Shenzhen

Abstract. We present a new framework to reconstruct holistic 3D indoor scenes including both room background and indoor objects from single-view images. Existing methods can only produce 3D shapes of indoor objects with limited geometry quality because of the heavy occlusion of indoor scenes. To solve this, we propose an instance-aligned implicit function (InstPIFu) for detailed object reconstruction. Combining with instance-aligned attention module, our method is empowered to decouple mixed local features toward the occluded instances. Additionally, unlike previous methods that simply represents the room background as a 3D bounding box, depth map or a set of planes, we recover the fine geometry of the background via implicit representation. Extensive experiments on the SUN RGB-D, Pix3D, 3D-FUTURE, and 3D-FRONT datasets demonstrate that our method outperforms existing approaches in both background and foreground object reconstruction. Our code and model will be made publicly available.

(a) Input image (b) Reconstructed scene (c) Total3D (d) Im3D (e) Ours

Fig. 1. Given a single indoor scene image, we reconstruct the holistic scene with detailed geometry, including the room background and indoor objects. From left to right: input image, the scene reconstructed by our method, results of Total3D [34], Im3D [56] and our method in a different camera pose.

⋆ Both authors contributed equally to this paper.
† **Corresponding Email**: hanxiaoguang@cuhk.edu.cn

1 Introduction

With the development of virtual reality (VR) and augmented reality (AR), the requirements for understanding and digitizing real-world 3D scenes are getting higher, especially for the indoor environment. If reconstructing the holistic indoor scene can be as simple as taking a picture using a mobile phone, we can efficiently generate a large scale of high-quality 3D content and further promote the development of VR and AR. Also, robots can better understand the real-world with the advance of single-view scene reconstruction. Hence, the problem of holistic indoor scene reconstruction from a single image has attracted considerable attention in recent years.

Early methods simplify this problem as estimating the room layout [15, 24, 28, 5, 40] and indoor objects [7, 17, 2] as 3D bounding boxes. However, such a coarse representation can only provide scene context information but cannot provide shape-level reconstruction. Mesh retrieval based approaches [20, 19, 18] improve the object shapes by substituting the 3D object boxes with meshes searched from a database. Due to the various categories and appearances of indoor objects, the size and diversity of the database directly influence the accuracy and time efficiency of these methods.

Inspired by learning-based shape reconstruction methods, voxel representation [25, 49, 22] is first applied to recover the 3D geometry of indoor scenes, but the shape quality is far from satisfactory due to the limited resolution. Mesh R-CNN [12] can reconstruct meshes for multiple instances from a single-view image, but lacks of scene understanding. Recently, Total3D[34] and Im3D[56] are proposed to reconstruct the 3D indoor scene from a single image, where the instance-level objects are represented in the form of explicit mesh and implicit surface, respectively. Although they have achieved state-of-the-art results on this task, they still have the following limitation. First, they often output shapes lacking details, due to the issue of limited training data and the use of global image feature for shape reconstruction. Second, the room layout in their methods is expressed as a simplified representation (*i.e.*, the 3D bounding box) which cannot recover backgrounds with complex geometries, like non-planar surfaces.

Recently, pixel-aligned implicit function (PIFu) has achieved promising results for detailed and generalizable 3D human reconstruction from a single image [42]. Motivated by the success of PIFu, we address the limitations of existing methods by introducing an *instance-aligned implicit function* (InstPIFu) for holistic and detailed indoor scene reconstruction from a single image. Note that pixel-aligned feature cannot be straightforwardly applied to indoor scene reconstruction, as objects (*e.g.*, sofa, chair, bed, and other furniture) are often occluded in a cluttered scene (see Fig. 1), such that the extracted local feature might contain mixed information of multiple objects. It is sub-optimal to directly use such a contaminated local feature for implicit surface reconstruction. To tackle this problem, we introduce an *instance-aligned attention module*, consisting of *attentional channel filtering*, and *spatial-guided supervision* strategies, to decouple the mixed local features for different instances in the overlapping regions.

Unlike previous methods that simply recover the room layout as a 3D bounding box [15, 24, 28, 5, 40, 34, 56], sparse depth [49] or room layout structure [57, 44, 55] without non-planar geometry, our implicit surface representation allows the detailed shape reconstruction of the room background (*e.g.*, floor, wall, and ceiling). Compared with existing approaches that encode the latent shape code with global image features [34, 56], the instance-aligned local features utilized in our encoder help alleviate the over-fitting problem and recover more detailed geometry of indoor objects. Extensive experiments on the SUN RGB-D, Pix3D, 3D-FUTURE, and 3D-FRONT datasets demonstrate the superiority of our method.

The key contributions of this paper are summarized as follows:

- We introduce a new pipeline to reconstruct the holistic and detailed 3D indoor scene from a single RGB image using implicit representation. To our best knowledge, this is the first system that uses pixel-aligned feature to recover the 3D indoor scene from a single view.
- We are the first to attempt to reconstruct the room background via implicit representation. Compared to previous methods that represent room layout as a 3D box, depth map or a set of planes, our method is capable to recover background with more complex geometries, like non-planar surfaces.
- We propose a new method, called InstPIFu, to use the instance-aligned feature, extracted by a novel instance-aligned attention module, for detailed indoor object reconstruction. Our method is more robust to object occlusion and has a better generalization ability on real-world datasets.
- Our method achieves state-of-the-art performance on both the synthetic and real-world indoor scene datasets.

2 Related Work

Single-view indoor scene reconstruction The long-standing problem of indoor scene reconstruction from a single image aims to construct the holistic 3D scene, which entails room layout estimation, object detection and pose estimation, as well as 3D shape reconstruction. Early works first recover the room layout as a 3D room bounding box [15, 24, 28, 5, 40]. Follow-up works make rapid progress toward object pose recovery [7, 17, 2], but still represent objects as 3D boxes without shape details.

To recover object shapes, some methods search for models with a similar appearance from a database [20, 19, 18]. However, the mismatch between objects in images and the database often leads to unsatisfactory results. Other methods [25, 49, 22] try to reconstruct the voxel representation for each object instance, but they are subjected to the problem of limited resolution. Mesh R-CNN [12] is capable to reconstruct meshes for multiple objects from a single-view image, but ignores scene understanding. To overcome the above limitations of previous solutions, Total3D [34] proposes an end-to-end system to jointly reconstruct room layout, object bounding boxes, and meshes from a single image. But its mesh generation network can only produce non-watertight mesh when handling shapes

with complex topology. The following Im3D [56] represents each object with the implicit surface function that can be converted to a watertight mesh via marching cube algorithm while preserving geometry details in the meantime. However, the state-of-the-art solution of Im3D [56] still suffers from shape over-fitting due to the problem of limited training data and the use of global image features for shape reconstruction.

Room background representation Early methods [15, 24, 28, 5, 40] simply recover the room background as a 3D bounding box, but room is usually not a cuboid. The state-of-the-art single-view indoor scene reconstruction methods [34, 56] are still using this representation for room background. [49] predicts the background via depth estimation, which recovers more details for background. However, the accuracy of background depth estimation is far from satisfactory because of the occlusion of foreground, *i.e.*, indoor objects. Recent works try to reconstruct the room layout structure [57, 44, 55] with the assumption that the background of the room (*e.g.*, floor, wall, and ceiling) is mainly composed of planes. Hence, only planar geometry can be recovered and nonplanar information is missed by these methods.

Learning-based 3D shape reconstruction Recent learning-based methods have adopted different surface representations for 3D shape reconstruction, such as voxel, mesh, point cloud, patches, primitives, and implicit surface.

Voxel-based methods [4, 26, 51, 41, 47, 53] benefit from 2D CNNs because of the regularity of the voxel representation, but suffer from the balance between resolution and efficiency. Mesh-based methods reconstruct the mesh of an object through deforming a template (*e.g.*, a unit sphere), but the topology of the obtained mesh is restricted [52, 13, 36, 21]. To modify the topology, some approaches learn to remove extra edges and vertices [36, 46, 34], which results in non-watertight meshes. Methods based on point cloud [9, 29, 23, 32], patches [13, 53], and primitives [48, 50, 38, 6] are adaptable to complex topology, but require post-processing to convert to structural representations. However, the post-processing is difficult to preserve the detailed geometry of the shape. Recently, implicit surface function [37, 3, 31, 54, 30] has been widely adopted as it can achieve detailed reconstruction for shape with an arbitrary topology and is easy to be converted to fine mesh.

Pixel-aligned image features Single-view implicit surface reconstruction methods often adopt an encoder-decoder pipeline and learn a latent code from the input image for shape recovery. For time and memory efficiency, global image feature [37, 3, 35, 30, 8] is often adopted, but it cannot recover the local detailed information existed in the input image. As a result, coarse results often occur in these approaches. Recently, pixel-aligned local image features have been demonstrated to recover complex geometries from a single view [42, 54].

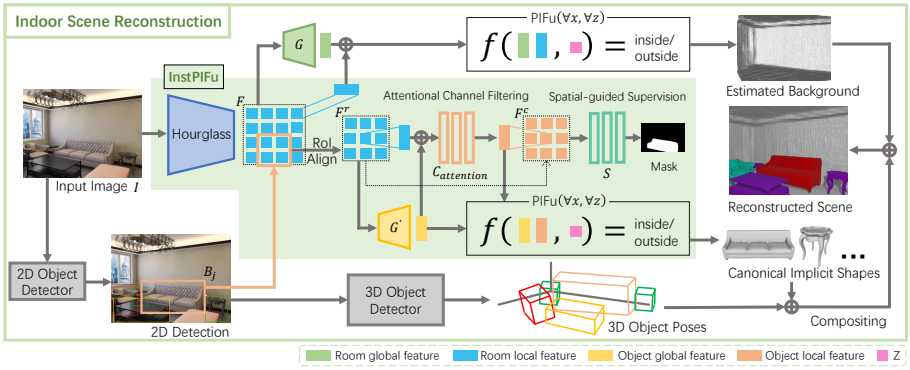


Fig. 2. Overview of the proposed InstPIFu. Given a single indoor scene image as input, our method simultaneously performs room background estimation, object detection and camera pose estimation, as well as detailed 3D object reconstruction.

3 Instance-aligned Implicit Representation

In this section, we first review the pixel-aligned implicit function (PIFu) and point out its limitation in dealing with the occluded object in the indoor scene. We then introduce our instance-aligned implicit function to perform better indoor object reconstruction where objects are often occluded in the cluttered scene.

3.1 Review of Pixel-aligned Implicit Modeling

Single-view scene reconstruction benefits from implicit representation [56], but the usage of global image features often causes coarse results. PIFu with pixel-aligned local features has been witnessed to recover detailed shapes in 3D human reconstruction [42].

A 3D surface can be defined by an implicit function as a level set of function f, *e.g.* $f(X) = 0$, where X is a 3D point. Similarly, a pixel-aligned implicit function f, represented by multi-layer perceptrons (MLPs), defines the surface as a level set of

$$f(F(x), z(X)) = s : s \in \mathbb{R}, \tag{1}$$

where $x = \pi(X)$ gives the 2D image projection point of X, $F(x) = g(I(x))$ is the local image feature at x extracted by a fully convolutional image encoder g, and $z(X)$ is the depth value in weak-perspective camera coordinate. We observe that adding the global image feature as an extra input helps in shape reconstruction. The adapted PIFu used in this work is defined as

$$f(F(x), F^G(I), z(X)) = s : s \in \mathbb{R}, \tag{2}$$

where $F^G(I)$ represents the global features of image I encoded by G.

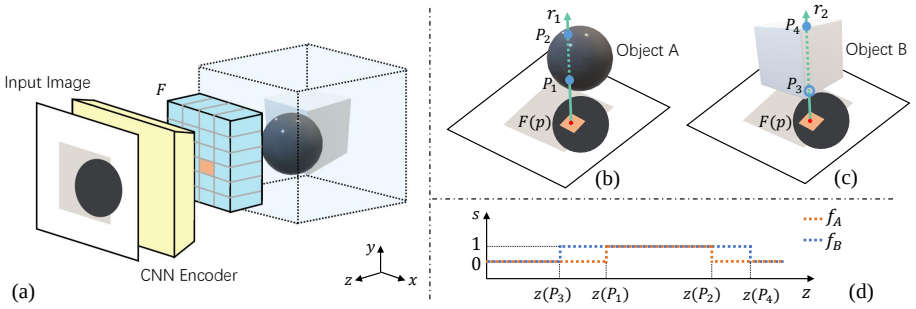


Fig. 3. Occlusion causes local feature ambiguity among different objects. (a) A scene contains two objects, and F is the extracted local feature from the image. (b)-(c) Object reconstruction in canonical coordinate system, where points along the rays are projected at p to sample local feature $F(p)$. (d) Variations of occupancy s with depth z along the ray r_1 and r_2 for f_A and f_B.

3.2 Limitation of Pixel-aligned Feature

Although PIFu demonstrates detailed reconstruction results in single human reconstruction, applying PIFu for indoor object reconstruction straightforwardly is not good, as it suffers a lot from the object occlusion that leads to feature ambiguity. Multiple 3D points belonging to different objects can be projected into similar 2D image location and get the same local image feature, such that the local feature will contain mixed information from different instances, which is not desirable for shape reconstruction.

As an example in Fig. 3 (a), a scene consists of a sphere A and and a cube B, where A occludes B in the captured image. Fig. 3 (b)-(c) show that when sampling pixel-aligned features for 3D points, points along the ray r_1 and r_2 (*e.g.*, P) are all projected at the point p in the overlapping region. This means that the same local feature $F(p)$ will be used to compute the occupancy value s for implicit function of A (f_A) and B (f_B), *i.e.*, $s = f_A(F(p), z(P))$ and $s = f_B(F(p), z(P))$. As PIFu implemented f_A and f_B using the same MLPs, adopting the same local feature $F(p)$ raises feature ambiguity in occupancy estimation for A and B. This is illustrated in Fig. 3 (d), where variations of s with z for f_A and f_B are apparently different. Note that here we simply represent the PIFu f as an ideal occupancy field where levels of points inside the object are 1, otherwise 0.

One possible solution might be adding the global feature of the instance as extra inputs to the shape decoder. But only using global features to tackle the ambiguity in occlusion region is not enough (see our ablation study). Because the local features still contain mixed information from different instances.

3.3 Instance-aligned Feature Concentration

To address the above limitation of PIFu, we propose InstPIFu, which adopts an instance-aligned attention module to disentangle the mixed feature information caused by object occlusion, for indoor object reconstruction. The proposed

instance-aligned attention module reduces the ambiguity of the local image feature by three sequential steps, *i.e.*, *RoI alignment*, *attentional channel filtering*, and *spatial-guided supervision* (see Fig. 2).

RoI alignment The first step is to extract instance-related features for each instance. A straight-forward solution is to extract features independently from the cropped image patch of each target instance. However, it is inefficient when there are multiple objects in a cluttered scene, and the useful scene contextual information will be ignored. Instead, we follow Mask R-CNN [14] to use RoI alignment for instance-related feature extraction. Given an image I and the 2D bounding box B_j of an instance j, we first crop out the corresponding local features of the region of interest (RoI) from the whole pixel-aligned feature map F and align them to F^r as [14]

$$F^r = \text{RoIAlign}(F, B_j). \tag{3}$$

Note that F^r has a fixed size of $W_r \times H_r$ for input feature maps with different shapes and the 2D bounding box B_j of object j is obtained by a Faster R-CNN detector [39]. We then extract a global instance feature for instance j as $G'(F^r)$, where G' is a global instance image encoder. The global instance feature will be used to compute the channel-wise attention for local feature filtering.

Attentional channel filtering Each local feature in the aligned RoI feature map F^r will be concatenated with the global instance feature $G'(F^r)$ as input for a channel-wise attention layer, similar to the Squeeze-and-Excitation block in [16] structurally, to generate an attention map with the same channel number of L_c as the local feature. This attention map will multiply with the local feature to filter out irrelevant feature by channel filtering to allow the updated local feature to concentrate on the target instance. This operation can be expressed as

$$F^c(x) = C_{\text{attention}}(F^r(x), G'(F^r)) \times F^r(x), \tag{4}$$

where $F^c(x)$ is the filtered local image feature at the 2D projection x of a 3D point X for instance j. Note that $F^r(x)$ in Eq. (4) adopts bilinear interpolation to access the features, and x in $F^r(x)$ should be shifted and scaled as well.

Spatial-guided supervision To better guide the learning of the channel filtering, we need a module that can encourage the filtered feature to focus more on the target instance. Thus, we exploit a spatial-guided supervision on the the filtered local feature map F^c that is the output of the channel-wise attention layer with the same shape as F^r. The Feature map F^c will be fed into a fully convolutional layer S to estimate a complete mask M for the target instance, *i.e.*, $M = S(F^c)$. This spatial-guided supervision can filter out irrelevant information out of the mask.

Instance-aligned implicit function Given the instance-aligned feature, we define InstPIFu f_o as

$$f_o(F^c(x), G'(F^r), z(X)) = s : s \in \mathbb{R}. \tag{5}$$

By applying the proposed instance-aligned attention module for decoupling the mixed local feature, compared with PIFu, the local feature used in our InstPIFu provides more discriminative information for accurate and detailed shape reconstruction. And this can be demonstrated by our ablation study.

4 Holistic Indoor Scene Reconstruction

Given a single image of an indoor scene, we aim to recover the holistic and detailed 3D scene in implicit representation (see Fig. 2). This problem is normally divided into several sub-tasks, including room background estimation, 3D object detection (pose estimation), as well as instance-level object reconstruction [34, 56]. We first process these three tasks individually and then perform scene compositing for holistic scene reconstruction. Note that our method recovers the room background with geometry details instead of just a simplified 3D bounding box.

4.1 Room Background Estimation

Room is usually not a cuboid. Thus, it is inappropriate to represent the room background as a 3D bounding box like [34, 56]. Depth map [49] is also not an ideal representation, because the accuracy of background depth estimation is heavily influenced by the occlusions of indoor objects in front of the background. Also, methods [44, 55] based on plane detection cannot recover small planes and non-planar background geometries. To address the above issues, we explore to use the implicit representation for room background reconstruction in this work.

The ground-truth room surface is represented as a 0.5 level set and then discretized to a 3D occupancy field:

$$f_r^*(X) = \begin{cases} 1, & \text{if } X \text{ is inside the room} \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Compared with indoor objects that have various styles and complicated geometries, the shape of the room background is much simpler. We find that applying the adapted PIFu (see Eq. (2)) which takes pixel-aligned features and global features for room background reconstruction already achieves good results. We train our room estimation PIFu f_r by minimizing the average of mean squared error (MSE):

$$\mathcal{L}_r = \frac{1}{n} \sum_{i=1}^{n} \left| f_r\left(F\left(x_i\right), G\left(F\right), z\left(X_i\right)\right) - f_r^*\left(X_i\right) \right|^2, \tag{7}$$

where n is the number of sample points, $X_i \in \mathbb{R}^3$ is a point in the camera coordinate system, $F(x) = g(I(x))$ is the local image feature located at x, $G(F)$ is the global image feature of the room background and F is the whole feature map produced by Hourglass network. The local and global image features are both from a stacked hourglass network [33], but an extra global encoder G is needed to encode the whole feature map F to the global feature. The obtained implicit room background can be easily converted to an explicit mesh via marching cube algorithm.

4.2 Indoor Object Reconstruction

As discussed in Sec. 3, due to the heavy occlusions between indoor objects, directly applying PIFu for instance reconstruction suffers from the problem of ambiguous local features. We adopt the proposed InstPIFu, which applies instance-aligned attention module for feature filtering, to reconstruct the indoor objects. We define the ground-truth surface of an indoor object as the room background (see Eq. (6)). The InstPIFu f_o is also trained by minimizing the average of MSE:

$$\mathcal{L}_o = \frac{1}{n} \sum_{i=1}^{n} \left| f_o \left(F^c \left(x_i \right), G'(F^r), z \left(X_i \right) \right) - f_o^* \left(X_i \right) \right|^2, \tag{8}$$

where $X_i \in \mathbb{R}^3$ is a point in the canonical coordinate system. Note that the projection from X_i to x_i is different from the original PIFu. Because X_i is in object coordinate system, extra camera and object poses are needed when projecting. We follow [56] to predict these parameters for projecting. The channel-wise attention layer is implemented as MLPs. During training, we add an extra instance mask loss for the instance-aligned attention module to enforce the feature to be constrained on the corresponding instance mask. The mask loss is simply implemented by the MSE between the predicted mask and the ground truth.

4.3 Scene Compositing

The room background is obtained in the camera coordinate system, while the objects are recovered in their canonical coordinate system to ease the learning of reconstructing indoor objects with various poses and scales. To embed objects into the scene together with the room background, the camera pose $\mathbf{R}(\beta, \gamma)$ and object bounding box parameters (δ, d, s, θ) are required. We use similar camera estimator and 3D object detector to predict above parameters as [34, 56]. Additionally, the Scene Graph Convolutional Network proposed in [56] is also used in our work to improve the performance of camera and object pose estimation. Note that we use perspective camera model.

5 Experiment

5.1 Experiment Setup

Datasets We conduct experiments on both synthetic and real datasets. The proposed pipeline is trained on 3D-FRONT [10] which is a large-scale repository

of synthetic indoor scenes, consisting of professionally designed rooms populated by 3D furniture models with high-quality geometry and texture in various styles. The furniture models come from 3D-FUTURE [11]. We use about 20K scene images for training and 6K for testing, where more than 16K objects from 3D-FUTURE are included. Following [34, 56], we also evaluate our method on real-world datasets: SUN RGB-D [43] and Pix3D [45].

Metrics We adopt the commonly used Chamfer distance (CD) to evaluate the background reconstruction, as it is difficult to compare our background results with layout Intersection over Union (IoU) [34, 56, 55, 44] (detailed reasons in Sec. 5.2). The reconstructed indoor objects are evaluated with CD and F-Score [52, 34, 56].

5.2 Evaluation on Room Background Estimation

We first evaluate the effectiveness of our room background estimation module. Layout IoU is a commonly used metric when comparing the room background. It is computed using the layout structure of the whole room. However, our method only reconstructs partial room background within the camera view. Hence, to compare our room background results with existing methods quantitatively, we firstly sample 10K points within the camera frustum from the reconstructed background in representations of bounding box [56], depth map [49, 1], plane sets [27] and our implicit surface, then compute CD with points on ground truth background. We choose to compare with PlaneRCNN [27] since it is popular and has decent performance in plane estimation. Because Factored3D [49] is based on depth estimation, we also compare it with Adabins [1] that is the state-of-the-art depth estimation approach. Quantitative comparisons in Tab. 1 shows the superiority of our method in detailed background recovery. Visual results of background reconstruction on 3D-FRONT and SUN RGB-D show that our method can recover the geometry details of the room background (see Fig. 5). More visual comparisons are given in the Supplementary Material.

Method	Factored3D [49]	Adabins [1]	Im3D [56]	PlaneRCNN [27]	Ours
CD on 3D-FRONT ↓	0.697	0.573	1.974	0.717	**0.481**

Table 1. Quantitative comparisons of room background estimation on 3D-FRONT.

5.3 Evaluation on Indoor Object Reconstruction

We compare our InstPIFu against the MGN of Total3D [34] and the LIEN of Im3D [56] on indoor object reconstruction. Quantitative and qualitative comparisons are shown on both 3D-FUTURE and Pix3D. Furthermore, we also train and test these object reconstruction networks on Pix3D with a non-overlapped

Fig. 4. Qualitative comparisons of indoor object reconstruction. From left to right of every quintuplet: (1) Input images and results from (2) MGN [34], (3) LDIF [56], (4) Ours, (5) Ground truth. The first two rows are compared on 3D-FUTURE, and the last two rows are on Pix3D. Note that results of the last row are generated by models trained and tested on non-overlapped split.

split to evaluate their generalization ability. CD is used to evaluate on the 10K points sampled from the reconstructed mesh after being aligned with the ground-truth using ICP. Note that results generated by InstPIFu and LIEN are in implicit representation which are converted to mesh using marching cube algorithm with a resolution of 256.

Evaluation on 3D-FUTURE Tab. 2 summarizes the quantitative results on 3D-FUTURE evaluated on 2000 indoor objects in 8 different categories. We use scene images in 3D-FRONT as the input for our InstPIFu, and cropped patches by ground-truth 2D bounding boxes (following [34, 56]) from every scene image as the input for MGN and LIEN. In these input images, object occlusions often occur. And thanks to the use of the instance-aligned feature, our method achieves the best on F-Score and shows decent results on CD (see Tab. 2). Although explicit methods like MGN achieve better CD loss as they directly optimize the CD loss during training, the reconstructed meshes lack details [30, 36, 54]. Also, MGN can not generate watertight mesh which is desired in object reconstruction. Fig. 4 (first two rows) shows that the results of our method have the most similar appearances to objects in the input images.

Comparison on Pix3D Quantitative results on Pix3D using the train/test split in [34] are shown in Tab. 3, where LIEN and MGN achieve better than ours. The major reason is that LIEN and MGN tend to be over-fitting on Pix3D which has only about 400 shapes. Because the split in [34] is based on different images, and all shapes in testing dataset also occur in training dataset. Also, the usage of pixel-aligned local feature makes our model achieve better generalization ability, but weaken the fitting performance. Nevertheless, our method still achieves comparable qualitative results (see the third row in Fig. 4).

Comparison of generalization To compare the generalization ability of the above three object reconstruction networks, we re-split Pix3D based on different shapes

Method	bed	chair	sofa	table	desk	nightstand	cabinet	bookshelf	mean ↓ / ↑
MGN [34]	**15.48** / 46.81	**11.67** / 57.49	8.72 / 64.61	**20.90** / 49.80	**17.59** / 46.82	17.11 / 47.91	13.13 / 54.18	10.21 / 54.55	**14.07** / 55.64
LIEN [56]	16.81 / 44.28	41.40 / 31.61	9.51 / 61.40	35.65 / 43.22	26.63 / 37.04	16.78 / 50.76	7.44 / 69.21	11.70 / 55.33	28.52 / 45.63
Ours	18.17 / **47.85**	14.06 / **59.08**	**7.66** / **67.60**	23.25 / **56.43**	33.33 / **48.49**	**11.73** / **57.14**	**6.04** / **73.32**	**8.03** / **66.13**	14.46 / **61.32**

Table 2. Quantitative comparisons of object reconstruction on 3D-FUTURE (CD / F-Score). The values of CD are in units of 10^{-3}.

Split in [34]	bed	bookcase	chair	desk	sofa	table	tool	wardrobe	misc	mean ↓ / ↑
MGN [34]	5.99 / **78.08**	6.56 / 62.98	**5.32** / **72.73**	**5.93** / 75.04	**3.36** / **79.64**	14.19 / 65.27	3.12 / **81.17**	3.83 / 85.51	26.93 / 46.76	6.84 / **73.18**
LIEN [56]	**4.11** / 65.26	**3.96** / 46.05	5.45 / 59.84	7.85 / **76.03**	5.61 / 64.02	**11.73** / **72.28**	**2.39** / 36.09	4.31 / 58.59	**24.65** / **57.50**	**6.72** / 63.96
Ours	9.52 / 59.47	4.38 / **73.25**	14.40 / 48.26	13.70 / 64.24	8.21 / 57.17	22.6 / 57.52	7.76 / 69.36	**3.67** / **87.36**	30.32 / 35.05	13.60 / 56.07
Non-overlapped Split	bed	bookcase	chair	desk	sofa	table	tool	wardrobe	misc	mean ↓ / ↑
MGN [34]	22.91 / 34.69	36.61 / 28.42	56.47 / **35.67**	33.95 / 34.90	9.27 / 51.15	81.19 / 17.05	94.70 / 57.16	10.43 / 52.04	137.5 / 10.41	44.32 / 36.20
LIEN [56]	11.88 / 37.13	29.61 / 15.51	40.01 / 25.70	65.36 / 26.01	10.54 / 49.71	146.13 / 21.16	29.63 / 5.85	4.88 / 59.46	144.06 / 11.04	51.31 / 31.45
Ours	**10.90** / **54.99**	**7.55** / **62.26**	**32.44** / 35.30	**22.09** / **47.30**	**8.13** / **56.54**	**45.82** / **37.51**	**10.29** / **64.24**	**1.29** / **94.62**	**47.31** / **27.03**	**24.65** / **45.62**

Table 3. Quantitative comparisons of object reconstruction on Pix3D with split in [34] and non-overlapped split.

(70% for training and 30% for testing), which ensures that all shapes in testing dataset have not been seen when training (non-overlapped split). Quantitative results are shown in Tab. 3, where our method achieves the best result due to the use of local image features. In contrast, MGN and LIEN suffer from over-fitting caused by global image features. Qualitative results in Fig. 4 (the last row) give the same conclusion, where objects reconstructed by MGN and LIEN are coarse shapes. More results are shown in the supplementary material.

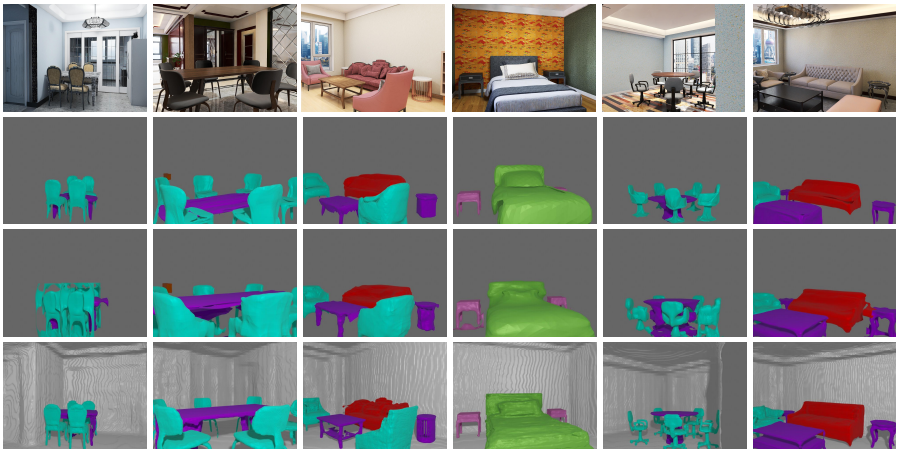
5.4 Qualitative Result of Holistic Scene Reconstruction

We compare our method with Total3D [34] and Im3D [56] in holistic indoor scene reconstruction on both 3D-FRONT [10] and SUN RGB-D [43] datasets. Qualitative comparisons shown in Fig. 5 demonstrate the superiority of our instance-aligned implicit representation. For fair comparison on SUN RGB-D, we first train the InstPIFu on 3D-FRONT and 3D-FUTURE and then finetune it on Pix3D. And we also use the predicted 3D object boxes by Im3D. Although our reconstructed scenes on SUN RGB-D may have some noisy patches due to the domain gap between the synthetic and the realistic datasets, the results are full of details in both the background and indoor objects, which reveals the good generalization ability of our method to some extend.

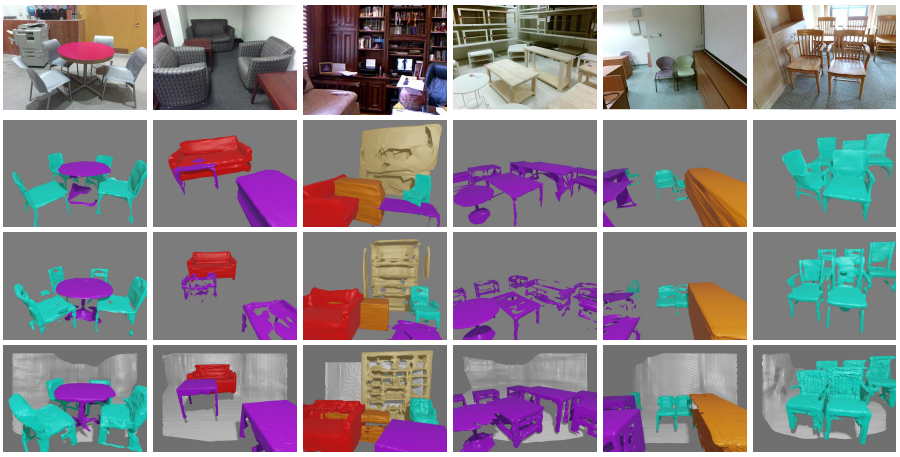
5.5 Ablation Study

To better study the effect of instance-aligned implicit representation for indoor object reconstruction, our method is ablated with five configurations:
- ***Baseline*:** only pixel-aligned feature is used in object reconstruction.
- $\boldsymbol{C_0}$**:** pixel-aligned feature + global instance feature.
- $\boldsymbol{C_1}$**:** $\boldsymbol{C_0}$ + attentional channel filtering.
- $\boldsymbol{C_2}$**:** $\boldsymbol{C_0}$ + spatial-guided supervision.
- ***Full*:** $\boldsymbol{C_0}$ + attentional channel filtering + spatial-guided supervision.

As the quantitative comparisons shown in Tab. 4, our ***Full*** model achieves the best results on metrics CD and F-Score, where we add the channel-wise attention together with the mask supervision to $\boldsymbol{C_0}$. If we remove the anyone of these two modules from ***Full***, that are $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$, CD and F-Score both



(a) Scene reconstruction results on 3D-FRONT



(b) Scene reconstruction results on SUN RGB-D

Fig. 5. Qualitative comparisons of holistic scene reconstruction. From the first row to the last: the input image, scene reconstruction results of Total3D, Im3D and ours. Note that the first four rows are compared on 3D-FRONT and the rest are on SUN RGB-D.

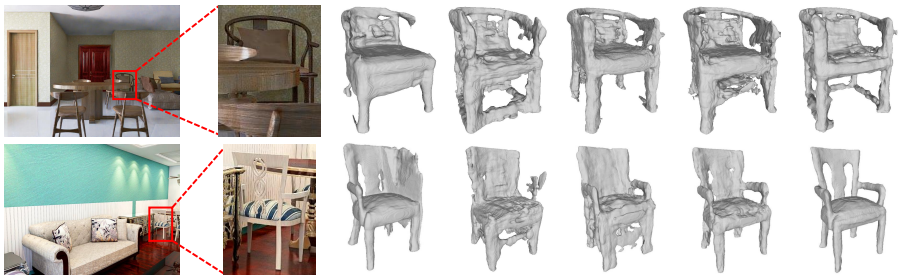


Fig. 6. Visual comparisons for ablation study. From left to right: the input image, results of ***Baseline***, $\boldsymbol{C_0}$, $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ and ***Full***.

Method	***Baseline***	C_0	C_1	C_2	***Full***
CD ↓	17.95	16.42(-1.53)	15.54(-2.41)	15.28(-2.67)	**14.46**(-3.49)
F-Score ↑	56.98	58.62(+1.64)	60.23(+3.25)	60.56(+3.58)	**61.32**(+4.34)

Table 4. Ablation study for the network architecture.

become worse. But C_1 and C_2 still perform better than C_0. This gives us the insight that both of channel-level filtering and spatial guidance help to decouple the feature ambiguity towards occluded objects. And the comparisons between the ***Baseline*** and C_0 show that concatenating global instance feature with pixel-aligned local feature is helpful for indoor object reconstruction. But from the comparisons of the whole table, we can see that only using global feature to tackle the ambiguity in occlusion region is not enough. Same conclusions can be drawn by the visual comparisons in Fig. 6.

6 Conclusion

We have introduced a new method based on implicit representation, called InstPIFu, for holistic and detailed 3D indoor scene reconstruction from a single image. To resolve the problem of ambiguous local features caused by object occlusions in an indoor scene, we proposed an instance-aligned attention module to effectively disentangle the mixed features for accurate instance shape reconstruction. Moreover, our method is the first to estimate the detailed room background via implicit representation, resulting in a more complete scene reconstruction. Extensive experiments on both synthetic and real datasets show that our method achieves state-of-the-art results for this problem.

Although our instance-aligned implicit function enables a more detailed and accurate indoor object reconstruction, the use of local feature makes the joint training of the 3D detection network and object reconstruction network not easy. Besides, real-world indoor scene datasets with high-quality 3D ground truth are scarce, and methods trained or finetuned with limited real data perform less well on real-world scenes compared with results on the synthetic scene (see Fig. 5). It would be interesting to explore how to take advantage of the existing large-scale and photo-realistic synthetic datasets for improving the generalization ability of the method.

Acknowledgement. The work was supported in part by the National Key R&D Program of China with grant No. 2018YFB1800800, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, and by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B12 12010001). It was also supported by NSFC-62172348, NSFC-61902334 and Shenzhen General Project (No. JCYJ20190814112007258). Thanks to the ITSO in CUHKSZ for their High-Performance Computing Services.

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021)
2. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. arXiv preprint arXiv:1909.01507 (2019)
3. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5939–5948 (2019)
4. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision. pp. 628–644. Springer (2016)
5. Dasgupta, S., Fang, K., Chen, K., Savarese, S.: Delay: Robust spatial layout estimation for cluttered indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 616–624 (2016)
6. Deprelle, T., Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: Learning elementary structures for 3d shape generation and matching. arXiv preprint arXiv:1908.04725 (2019)
7. Du, Y., Liu, Z., Basevi, H., Leonardis, A., Freeman, B., Tenenbaum, J., Wu, J.: Learning to exploit stability for 3d scene parsing. In: Advances in Neural Information Processing Systems. pp. 1726–1736 (2018)
8. Dupont, E., Martin, M.B., Colburn, A., Sankar, A., Susskind, J., Shan, Q.: Equivariant neural rendering. In: International Conference on Machine Learning. pp. 2761–2770. PMLR (2020)
9. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
10. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021)
11. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision pp. 1–25 (2021)
12. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. arXiv preprint arXiv:1906.02739 (2019)
13. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
15. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: 2009 IEEE 12th international conference on computer vision. pp. 1849–1856. IEEE (2009)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)

17. Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N., Zhu, S.C.: Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In: Advances in Neural Information Processing Systems. pp. 207–218 (2018)
18. Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, S.C.: Holistic 3d scene parsing and reconstruction from a single rgb image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 187–203 (2018)
19. Hueting, M., Reddy, P., Kim, V., Yumer, E., Carr, N., Mitra, N.: Seethrough: finding chairs in heavily occluded indoor scene images. arXiv preprint arXiv:1710.10473 (2017)
20. Izadinia, H., Shan, Q., Seitz, S.M.: Im2cad. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5134–5143 (2017)
21. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3907–3916 (2018)
22. Kulkarni, N., Misra, I., Tulsiani, S., Gupta, A.: 3d-relnet: Joint object and relational network for 3d prediction. International Conference on Computer Vision (ICCV) (2019)
23. Kurenkov, A., Ji, J., Garg, A., Mehta, V., Gwak, J., Choy, C., Savarese, S.: Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 858–866. IEEE (2018)
24. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2136–2143. IEEE (2009)
25. Li, L., Khan, S., Barnes, N.: Silhouette-assisted 3d object instance reconstruction from a cluttered scene. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
26. Liao, Y., Donne, S., Geiger, A.: Deep marching cubes: Learning explicit surface representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2916–2925 (2018)
27. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4450–4459 (2019)
28. Mallya, A., Lazebnik, S.: Learning informative edge maps for indoor scene layout prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 936–944 (2015)
29. Mandikal, P., KL, N., Venkatesh Babu, R.: 3d-psrnet: Part segmented 3d point cloud reconstruction from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
30. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
31. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Deep level sets: Implicit surface representations for 3d shape inference. arXiv preprint arXiv:1901.06802 (2019)
32. Navaneet, K., Mandikal, P., Agarwal, M., Babu, R.V.: Capnet: Continuous approximation projection for 3d point cloud reconstruction using 2d supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8819–8826 (2019)

33. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
34. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 55–64 (2020)
35. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020)
36. Pan, J., Han, X., Chen, W., Tang, J., Jia, K.: Deep mesh reconstruction from single rgb images via topology modification networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9964–9973 (2019)
37. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. arXiv preprint arXiv:1901.05103 (2019)
38. Paschalidou, D., Ulusoy, A.O., Geiger, A.: Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10344–10353 (2019)
39. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
40. Ren, Y., Li, S., Chen, C., Kuo, C.C.J.: A coarse-to-fine indoor layout estimation (cfile) method. In: Asian Conference on Computer Vision. pp. 36–51. Springer (2016)
41. Riegler, G., Osman Ulusoy, A., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3577–3586 (2017)
42. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)
43. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
44. Stekovic, S., Hampali, S., Rad, M., Sarkar, S.D., Fraundorfer, F., Lepetit, V.: General 3d room layout from a single view by render-and-compare. In: European Conference on Computer Vision. pp. 187–203. Springer (2020)
45. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2974–2983 (2018)
46. Tang, J., Han, X., Pan, J., Jia, K., Tong, X.: A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4541–4550 (2019)
47. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2088–2096 (2017)

48. Tian, Y., Luo, A., Sun, X., Ellis, K., Freeman, W.T., Tenenbaum, J.B., Wu, J.: Learning to infer and execute 3d shape programs. arXiv preprint arXiv:1901.02875 (2019)
49. Tulsiani, S., Gupta, S., Fouhey, D.F., Efros, A.A., Malik, J.: Factoring shape, pose, and layout from the 2d image of a 3d scene. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 302–310 (2018)
50. Tulsiani, S., Su, H., Guibas, L.J., Efros, A.A., Malik, J.: Learning shape abstractions by assembling volumetric primitives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2635–2643 (2017)
51. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3818–3827 (2019)
52. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–67 (2018)
53. Wang, P.S., Sun, C.Y., Liu, Y., Tong, X.: Adaptive o-cnn: a patch-based deep representation of 3d shapes. In: SIGGRAPH Asia 2018 Technical Papers. p. 217. ACM (2018)
54. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. arXiv preprint arXiv:1905.10711 (2019)
55. Yang, C., Zheng, J., Dai, X., Tang, R., Ma, Y., Yuan, X.: Learning to reconstruct 3d non-cuboid room layout from a single rgb image. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2534–2543 (2022)
56. Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S.: Holistic 3d scene understanding from a single image with implicit representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8833–8842 (2021)
57. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: Layoutnet: Reconstructing the 3d room layout from a single rgb image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2051–2059 (2018)