ISPRS Journal of Photogrammetry and Remote Sensing 186 (2022) 301–314

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

GeoRec: Geometry-enhanced semantic 3D reconstruction of RGB-D indoor scenes

Linxi Huan [a], Xianwei Zheng [a,c,*], Jianya Gong [a,b]

[a] The State Key Lab. LIESMARS, Wuhan University, Wuhan, PR China
[b] School of Remote Sensing and Engineering, Wuhan University, Wuhan, PR China
[c] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shen Zhen, PR China

ARTICLE INFO

Keywords:
Indoor 3D modeling
Semantic reconstruction
Deep learning
RGB-D scene

ABSTRACT

Semantic indoor 3D modeling with multi-task deep neural networks is an efficient and low-cost way for reconstructing an indoor scene with geometrically complete room structure and semantic 3D individuals. Challenged by the complexity and clutter of indoor scenarios, the semantic reconstruction quality of current methods is still limited by the insufficient exploration and learning of 3D geometry information. To this end, this paper proposes an end-to-end multi-task neural network for geometry-enhanced semantic 3D reconstruction of RGB-D indoor scenes (termed as GeoRec). In the proposed GeoRec, we build a geometry extractor that can effectively learn geometry-enhanced feature representation from depth data, to improve the estimation accuracy of layout, camera pose and 3D object bounding boxes. We also introduce a novel object mesh generator that strengthens the reconstruction robustness of GeoRec to indoor occlusion with geometry-enhanced implicit shape embedding. With the parsed scene semantics and geometries, the proposed GeoRec reconstructs an indoor scene by placing reconstructed object mesh models with 3D object detection results in the estimated layout cuboid. Extensive experiments conducted on two benchmark datasets show that the proposed GeoRec yields outstanding performance with 5.19×10^{-3} mean chamfer distance error for object reconstruction on the challenging Pix3D dataset, 70.45% mAP for 3D object detection and 77.1% 3D mIoU for layout estimation on the commonly-used SUN RGB-D dataset. Especially, the mesh reconstruction sub-network of GeoRec trained on Pix3D can be directly transferred to SUN RGB-D without any fine-tuning, manifesting a high generalization ability.

1. Introduction

Semantic-aware indoor 3D reconstruction is a comprehensive task that requires the recovery of the geometric and semantic context of an indoor scenario. An indoor semantic 3D model describes the basic 3D geometry of indoor space with rich scene knowledge, which brings benefits to a variety of location-based and intelligent indoor applications. The applications can range from indoor localization and intelligent navigation (Taira et al., 2021; Yang et al., 2021), 3D object retrieval and tracking (Zhou et al., 2018; Wald et al., 2019), augmented reality (Murez et al., 2020), interior design (Zhang et al., 2020b; Wang et al., 2020c), to indoor GIS analysis (Kang et al., 2020). However, semantic 3D indoor modeling remains an open challenge, due to the complex structure, high occlusion, and variability of indoor spaces. Technologies for outdoor 3D modeling, albeit extensively developed, still hardly satisfy the requirement of indoor applications.

For indoor semantic reconstruction, a general way is to build a 3D model with point cloud acquired by dense image matching or LiDAR observation (Izadi et al., 2011; Wang et al., 2020a; Wang et al., 2020b), and parse semantics subsequently based on the geometry of the 3D model (Koppula et al., 2011; Wang et al., 2018; Li et al., 2020b; Lin et al., 2021). Despite high accuracy for well observed regions, guaranteeing the completeness of the recovered 3D geometry remains hard in terms of data collection and traditional reconstruction procedure, which is disadvantageous for the following semantic parsing and individual object model recovery. Specifically, missing observation is inevitable due to high indoor occlusion, and repeat measurement is required for compensation, which can result in a high data collection cost when facing large-scale and frequently changed scenes. Even though with high-quality observation data, the traditional reconstruction system (*e.g.,* SfM) can also pose a high risk of failure in the regions with poor texture, repetitive structure and high reflection.

* Corresponding author.
E-mail address: zhengxw@whu.edu.cn (X. Zheng).

https://doi.org/10.1016/j.isprsjprs.2022.02.014
Received 31 July 2021; Received in revised form 7 February 2022; Accepted 18 February 2022
Available online 3 March 2022

Recently, increasing interest has been witnessed in a semantic-assisted reconstruction way. A semantic-assisted reconstruction system extracts 3D geometric properties and semantics from RGB/RGB-D data, and recovers an indoor semantic 3D model by combining the 3D room box and individual 3D object shapes that are restored with the extracted geometric and semantic clues. Complementary to the general solution, the semantic-assisted reconstruction provides a flexible and efficient way to reconstruct complete 3D models of indoor spaces/objects with semantics, and can separately reconstruct 3D room structure and object meshes with less requirement on the amount and density of observed data. A typical way for semantic-assisted indoor reconstruction is to search and align CAD exemplars with object parsing results (Gupta et al., 2015; Izadinia et al., 2017; Huang et al., 2018b; Avetisyan et al., 2019; Avetisyan et al., 2020). However, the shape retrieval step restricts the accuracy and efficiency of existing CAD-based methods to the size and diversity of the off-line CAD database, which is difficult to develop with various scenes (Zhang et al., 2021).

With the promising advances in 3D reconstruction with single color image brought by deep learning (Groueix et al., 2018; Xu et al., 2019; Genova et al., 2020), some researchers got rid of the pre-defined CAD pool by introducing reconstruction networks for end-to-end holistic scene recovery. Existing end-to-end methods usually realized indoor semantic reconstruction with a single RGB image and 2D object detection results. Some methods model scene layout and object shape via voxel occupancy grid, and integrate these independent volumetric models with predicted object poses (Tulsiani et al., 2018; Kulkarni et al., 2019; Li et al., 2020a), while some recent works predicted mesh representations for object modeling by performing topology modification on a template or learninga deep implicit function (DIF), and then arranged the reconstructed 3D shapes with inferred object bounding boxes and poses in the estimated 3D layout cuboid (Nie et al., 2020; Zhang et al., 2021). However, the voxel-grid representation is generally troubled by limited resolution, and rapidly expanding computational cost for high reconstruction quality; while current template- or DIF-based methods suffer from difficulties in modeling object with holes or heavily occluded/truncated instances. Despite the advances achieved with a single color image, the aforementioned RGB-based methods are generally not sufficiently robust to the clutter and complexity of indoor scenarios due to the ambiguity in depth, and the scene parsing accuracy of these methods is challenged by the heavy indoor occlusion and the large variation in object pose and scale. To excavate reliable 3D geometry information for scene understanding and reconstruction, numerous efforts were made on the exploration of depth information from RGB-D data (Qi et al., 2019; Zhang et al., 2020a; Avetisyan et al., 2020). Nevertheless, these studies either focused on a single vision task that is far from holistic indoor semantic reconstruction; or depended on densely captured RGB-D scans and an off-line CAD model dataset, which can impose burdens on computational efficiency and resources. It therefore remains an open challenge to effectively learn robust and representative geometric features from a single-view RGB-D image, for accurate geometric property parsing and occlusion-robust object reconstruction.

Based on the above observations, we propose a geometry-enhanced multi-task learning network for robust end-to-end semantic 3D reconstruction (GeoRec) of RGB-D scenes. The proposed GeoRec includes three main modules (sub-networks) for three sub-tasks, a layout estimator (GeoLE), an object detector (GeoOD) and a mesh generator (GeoMesh). GeoLE estimates camera pose and extracts layout structure properties to form the 3D room space; GeoOD detects 3D bounding boxes of object instances with poses; and GeoMesh reconstructs the meshes of 3D object instances and arranges them to corresponding locations according to the detected 3D object bounding boxes. The three modules coupled to accomplish the complete semantic 3D reconstruction of indoor scenes. To overcome the limitations of existing multi-task learning networks, our key idea is to learn rich geometry-enhanced features from RGB-D images for accurately capturing 3D geometric properties, and strengthening the robustness of instance-level modeling. For effective exploitation and utilization of the geometric information contained in depth modality, we build a novel geometry extractor that is capable of learning local and global geometric context from down-sampled point clouds generated from depth data, and adopt the geometry extractor for the three sub-tasks in semantic reconstruction. Specifically, the geometry extractor assists in inferring 3D layout cuboid and camera pose directly from down-sampled scene point clouds, and is fed with the point clouds restricted by the 2D object boxes derived from RGB images for detecting object geometric attributes. To learn occlusion-robust DIF 3D representation for object mesh modeling, GeoMesh is designed to learn geometry-enhanced implicit shape embedding that integrates the RGB-based visual clues and the global object structure information encoded by the geometry extractor.

We validated the proposed GeoRec on two commonly-used benchmarks, Pix3D (Sun et al., 2018) and SUN RGB-D datasets (Song et al., 2015). As the two datasets contain different types of ground-truth for different sub-tasks, we independently trained and verified the three modules on different datasets. In particular, due to the absence of 3D mesh gourd-truth in SUN RGB-D dataset, we directly transferred the GeoMesh trained on Pix3D to SUN RGB-D dataset for mesh reconstruction, the results of which manifests a very high generalization ability of GeoMesh. In general, the main contributions of our work can be summarized as follows.

- We design a geometry-enhanced multi-task learning network to jointly perform layout estimation, 3D object detection and instance-level reconstruction with an RGB-D image for high-quality indoor semantic reconstruction.
- We build a powerful geometry extractor that can encode rich and robust geometric features from depth data, to accurately estimate the room layout, the camera pose, and the 3D object bounding box.
- For modeling 3D object shape with robustness to occlusion, we design a novel mesh generator (GeoMesh) to learn occlusion-robust deep implicit function (DIF) representation with the global geometric structure information of objects.
- The GeoRec was validated on Pix3D and SUN RGB-D datasets. The GeoRec outperforms existing methods with 5.19×10^{-3} chamfer distance error for object reconstruction on Pix3D, 77.1% 3D mIoU for layout estimation, and 70.45% mAP for 3D object detection on SUN RGB-D.

The remainder of this paper is organized as follows. Section 2 presents a brief review of related works, and Section 3 elaborates the components of the proposed GeoRec. Section 4 provides the experimental results and ablation studies, while Section 5 finally draws the conclusions.

2. Related work

Automatic 3D scene reconstruction has been studied with a long history that can be at least traced back to pioneer work in 1963 (Roberts, 1963). Existing studies for indoor semantic reconstruction in general parse semantics based on 3D models restored by geometry reconstruction system, which has been demonstrated effective for outdoor environments. For instance, with the point clouds generated from dense image matching or LiDAR surveying, traditional hand-crafted 3D descriptors can be designed for 3D semantic object retrieval (Johnson and Hebert, 1999; Frome et al., 2004; Dong et al., 2017), while class-aware bounding boxes can be predicted for 3D object detection which group points that belong to the same instance (Qi et al., 2018; Gong et al., 2020). Semantic segmentation is also a popular way for acquiring semantics of 3D objects by assigning each 3D point with a category label (Qi et al., 2017; Landrieu and Simonovsky, 2018). Instead of parsing semantics on the basis of 3D models, a few researchers attempted mapping 2D semantic labels from images to restored 3D models, which

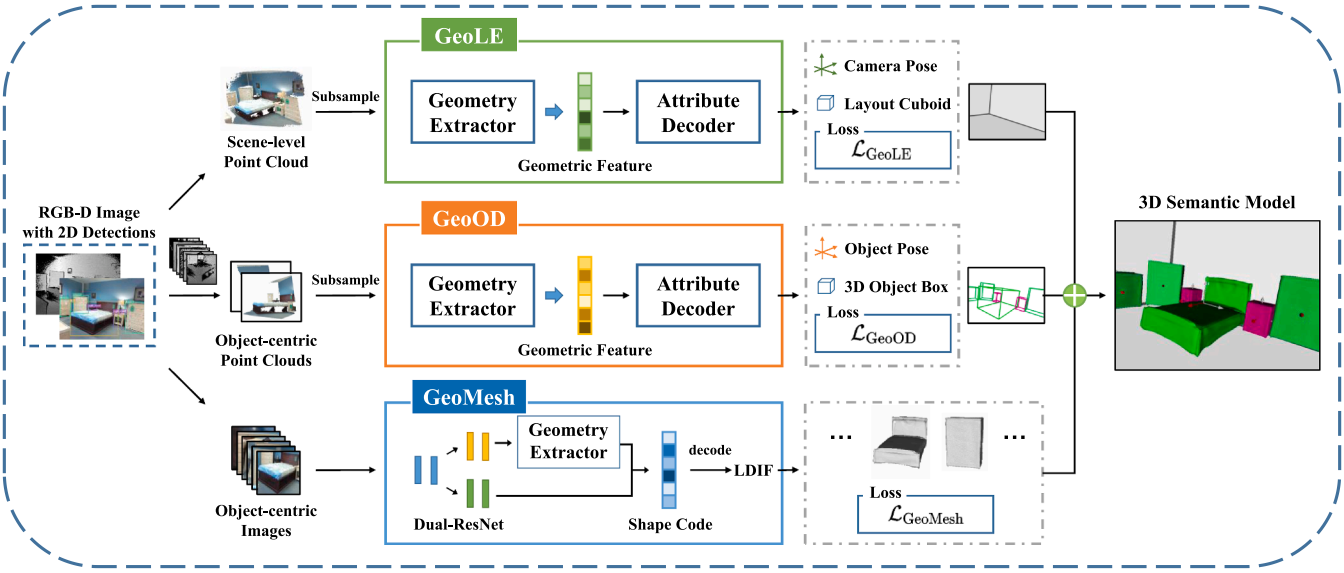

Fig. 1. Workflow of the proposed GeoRec. GeoLE is the geometry-enhanced layout estimator that derives the layout and camera pose from sub-sampled scene point clouds (generated from depth data), while GeoOD refers to the geometry-enhanced object detector that infers 3D object bounding boxes with object poses. The GeoMesh module restores 3D mesh representations for target instances with object-centric images produced by cropping the original RGB image with 2D detection results. The GeoRec finally reconstructs the indoor scene by integrating the 3D object shapes and the room layout cuboid with the parsed 3D geometric attributes of objects during joint inference.

circumvents the computation burden of direct operation on 3D data (Zhang et al., 2018; Dai and Nießner, 2018; Murez et al., 2020). However, geometry-based semantic parsing and 2D-3D semantic mapping are highly related to the reconstruction quality, which is often hampered by the geometry deficiency caused by issues such as occlusion, textureless regions and repetitive structures. The requirement for densely captured images or expensive professional laser devices also restricts the application of the aforementioned methods for low-cost and large-scale indoor semantic reconstruction.

With the development of many independent vision tasks, such as layout estimation and object detection (Lin et al., 2013; Lee et al., 2017; Qi et al., 2020), increasing attention from academia and industry has focused on building indoor semantic 3D models with monocular images or low-cost RGB-D data in a semantic-assisted way. Different from the traditional semantic methods that either depended semantic parsing on an already-built scene geometry model, or combined separately recovered scene geometry and semantics via mapping, a semantic-assisted reconstruction system recovers a room with learned geometric properties and scene semantics, such as layout cuboid and class-aware 3D bounding boxes.

With geometric attributes and semantics extracted from RGB/RGB-D images, one way for recovering the indoor semantic model is to align a CAD shape with the detected geometric and semantic information (Aubry et al., 2014; Bansal et al., 2016). These CAD-based methods modeled indoor objects with appearance-similar CAD shapes and placed the retrieved CAD models in a layout cuboid with alignment to the scene depicted in images. Align3D (Gupta et al., 2015) coarsely estimated object poses with instance semantic segmentation results and aligned prototypical models to the scene with the coarse poses and the inferred pixel support. IM2CAD (Izadinia et al., 2017) fit matched CAD model to a hypothesized room 3D box with information provided by object proposals. HoPR (Huang et al., 2018b) initialized a 3D indoor semantic representation with estimated layout and 2D object detections and refined the initial 3D model with holistic scene grammar in an analysis-by-synthesis manner. CooP (Huang et al., 2018a) combined a global geometry network (GGN) and a local object network (LON) for end-to-end scene recovery, where the GGN estimated a 3D layout cuboid and the related camera pose, while the LON learned 3D object bounding boxes along with object poses. Joint3D (Geiger and Wang, 2015) leverages the precise geometry of CAD models via inverse graphics for realizing holistic 3D scene understanding. Compared with HoPR, CooP and Joint3D that used monocular RGB/RGB-D images with 2D detection results, Scan2CAD (Avetisyan et al., 2019) and SceneCAD (Avetisyan et al., 2020) predicted a globally consistent CAD-based scene representation with the RGB-D scans derived from a long video sequence.

In contrast to the CAD-based methods, recently proposed approaches circumvented the reliance on a large off-line CAD library and the CAD retrieval problem by adopting volumetric and mesh 3D representations, which can be directly learned by deep neural networks (Groueix et al., 2018; Park et al., 2019; Genova et al., 2020). Factored3D (Tulsiani et al., 2018) is a typical voxel-based semantic reconstruction system that inferred the volumetric representations of scene layout and object shapes of a scene, and set the object models in the layout with predicted object poses. Inheriting the spirit of Factored3D that represented a room with an occupancy grid, Silhouette3D (Li et al., 2019) included learnable silhouette for improving object shape prediction, while 3D-RelNet (Kulkarni et al., 2019) incorporated pairwise relations between indoor instances for inferring the shape and pose of objects. As volumetric 3D representation often faces issues of low resolution and high computation burden, Total3D (Nie et al., 2020) started to build 3D object models for indoor semantic reconstruction by deforming the triangulation of a sphere template, and a subsequent work, namely Im3D (Zhang et al., 2021), further improved the reconstruction quality by learning DIF 3D representations with implicit embedding. However, the template-assisted topology deformation used in Total3D is weak in modeling objects with genus larger than zero; while the object shape inference of Im3D can completely fail when facing highly occluded or truncated objects, because of the incapacity of extracting global geometric information of object structure. Instead of focusing on single-view data, RevealNet (Hou et al., 2020) and RfD-Net (Nie et al., 2021) studied indoor semantic reconstruction with scans generated from RGB-D video sequences by realizing two sub-tasks including 3D object detection and point-based object mesh prediction.

Overall, single-view indoor semantic reconstruction is a cheap way to build indoor 3D models and also an efficient solution to deal with the frequently changed indoor spaces. However, despite the single-view indoor semantic reconstruction has been greatly advanced with many recent works (Tulsiani et al., 2018; Nie et al., 2020; Zhang et al., 2021),

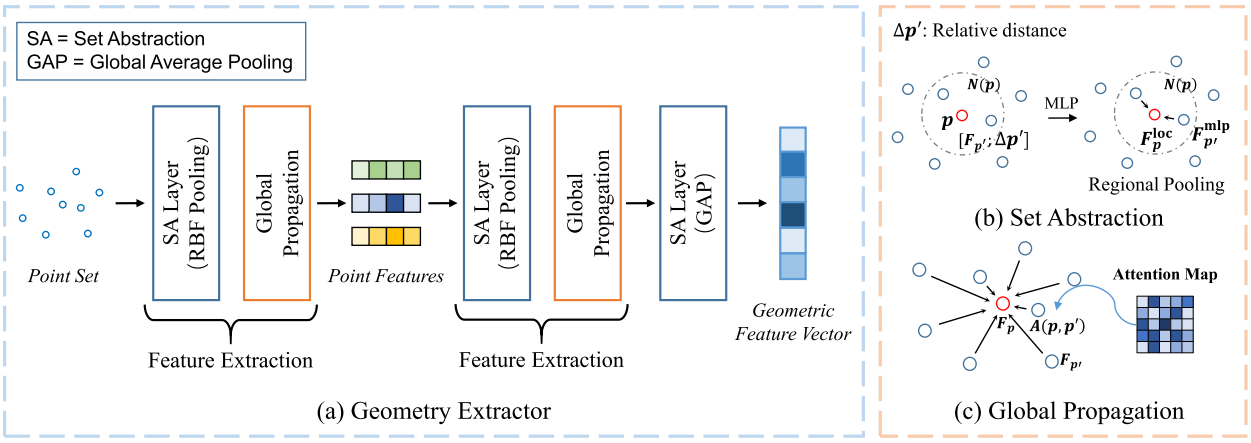

Fig. 2. Architecture and details of the geometry extractor. (a) Workflow of the proposed geometry extractor. (b) Mechanism of the set abstraction layer. $N(p)$ and F_p denote the neighbor point set and the feature of point p, respectively. $F_{p'}$ is a point feature that belongs to a point in $N(p)$ and $F_{p'}^{mlp}$ refers to the point feature learned by a multi-layer perceptron (MLP) with local relative distance clues. (c) The global feature propagation provided by the attention layer of a feature extraction module. $A(p, p')$ denotes the relation weight for the points features of p and p' in the attention map.

the depth ambiguity remains a problem for current methods. Considering that depth data can be easily captured along with RGB images in indoor scenarios via low-cost depth camera, we are motivated to study an effective exploration of the depth data for improving the performance of single-view indoor semantic reconstruction. Therefore, following the pipeline of Total3D and Im3D, we designed the GeoRec to perform indoor semantic reconstruction with a single-view RGB-D image under a multi-task framework, which simultaneously realizes the estimation of 3D layout and camera pose, the detection of objects and the reconstruction of object instances. Although there exist methods that utilizes RGB-D data for indoor semantic reconstruction, they generally rely on RGB-D scans derived from RGB-D video sequences or use the RGB-D data for CAD model alignment (Gupta et al., 2015; Hou et al., 2020; Nie et al., 2021). In contrast, our method explores single-view RGB-D data for indoor semantic segmentation, without the reliance on off-line CAD dataset and the heavy burdens of collecting and processing RGB-D sequence for scan generation in practice.

3. Methodology

In this section, we presents the overall pipeline of the proposed GeoRec for end-to-end indoor semantic reconstruction, and then elaborate the mechanism of our proposed geometry extractor that captures representative geometry features for all sub-tasks in GeoRec, including geometry property parsing and object shape recovery. On the basis of the geometry extractor, we subsequently introduce the architectures of a layout estimator, a 3D object detector, and especially, a novel object mesh generator that plays a key role in recovering the final indoor semantic 3D model.

3.1. Overview

In line with Total3D and Im3D, the proposed GeoRec achieves indoor semantic reconstruction by inferring the 3D geometric properties of layout and objects, and modeling object 3D shapes with class-aware 2D bounding boxes that are yielded by a 2D detector. Different from previous works, the GeoRec explores the effective utilization of a RGB-D image rather than a RGB image, to enhance the robustness of semantic reconstruction with rich geometric clues.

As depicted in Fig. 1, the proposed GeoRec comprises the following three main modules: a layout estimator (GeoLE), an object detector (GeoOD), and a mesh generator (GeoMesh). The GeoRec restores the final scene semantic model by jointly inferring and integrating the parsed geometric attributes and the estimated object meshes. Specifically, the layout estimator GeoLE is fed with sparse sub-sampled scene-level point clouds generated from depth data, and yields camera pose and layout structure properties, which can form the 3D room space. The object detector GeoOD encodes the geometry of down-sampled point clouds derived from depth images cropped by the 2D object bounding box to reduce the search space for 3D object detection (Qi et al., 2018), and decodes the geometric embedding into the parameters related to object poses and 3D bounding boxes. The mesh generator GeoMesh serves for instance-level 3D individuals reconstruction with object-centric images.

In the GeoRec method, a geometry extractor is constructed and applied for capturing rich geometric features from point cloud data, and a new mesh generator is designed to improve the reconstruction robustness to occlusion. Respectively, the GeoLE and GeoOD modules are formed with a geometry extractor as the backbone and a geometry attribute decoder that consists of several multi-layer perceptrons (MLPs), to parse 3D geometric properties with depth information. The novel object-wise reconstruction module (GeoMesh) is built with a geometry-enhanced shape encoder and a shape decoder. The geometry-enhanced encoder is composed of a Dual-ResNet backbone and a geometry extractor to learn latent shape embedding enhanced with global object structure information, which strengthens the completeness of restored 3D object shapes. The shape decoder then decodes the geometry-enhanced latent shape codes into DIF representations and recovers object meshes via the marching cube algorithm (Lorensen and Cline, 1987).

3.2. Geometry extractor

Indoor semantic reconstruction is a comprehensive problem that consists of multiple 3D parsing targets, and it will be thus impractical to assemble several computationally-expensive models for achieving different sub-tasks. In this case, we constructed a simple yet effective geometry extractor to learn rich and representative geometric features from sub-sampled depth-derived point clouds for sub-tasks, in which case the computation burden is reduced in terms of model complexity and data density. In this case, the layout estimator, the object detector and the mesh generator can not only benefit from the 3D geometry representations learned by the geometry extractor, but also work jointly to perform the end-to-end indoor semantic reconstruction.

The multiple vision tasks for indoor semantic reconstruction rely on geometric information of different scales to various degrees. For illustration, encapsulating an entire instance with precise a 3D bounding box may benefit more from global context, while estimating the pose of an

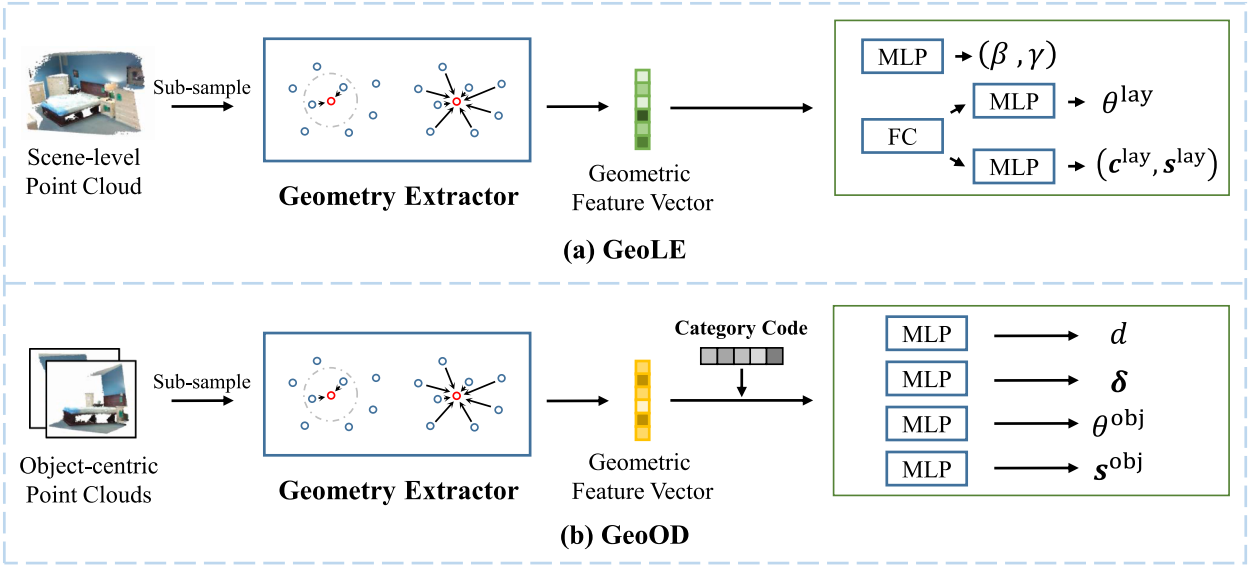

Fig. 3. Architectures of (a) GeoLE and (b) GeoOD. GeoLE or GeoOD is built on a geometry extractor to learn geometry feature vector, which is decoded by multi-layer perceptrons (MLPs) for geometric property parsing. GeoLE infers the pitch (β) and roll (γ) angles for the camera pose, and the layout-related parameters ($\boldsymbol{c}^{lay}$,$\boldsymbol{s}^{lay}$,θ^{lay}). GeoOD predicts ($\boldsymbol{\delta}, d, \boldsymbol{s}^{obj}$) and θ^{obj} for 3d object bounding boxes and object poses, respectively. One MLP consists of two fully connected convolution layers with a ReLU activation and a Dropout operation.

object may need more local shape clues. Therefore, the geometry extractor is expected to learn both the local and global geometric context for supporting different parsing targets. For this purpose, the geometry extractor, as displayed in Fig. 2(a), was built on the basis of the feature extraction module. This feature extraction module first performs local feature grouping with a set abstraction layer (Qi et al., 2017) and then globally propagates information across points to encode global geometric context.

In the local geometry embedding process described in Fig. 2(b), the set abstraction layer first selects centroids to divide the input into overlapped point sets. For each point in a given point set, the relative distance to the centroid is then embedded into the corresponding point feature with a shared multi-layer perceptron, and all the point features in the same point set are finally merged via regional pooling. Radial basis function (RBF) kernels are used for a weighted regional pooling considering that points closer to a centroid should contribute more to the shape around the centroid. Let p and F_p respectively denote a point and the related feature, while $N(p)$ is the set of points lying in the neighborhood of p, the weighted regional pooling can be formulated as Eq. (1),

$$F_p^{\text{loc}} = \sum_{p' \in N(p)} w_{p'} F_{p'}^{\text{mlp}} \Big/ |N(p)|, \quad w_{p'} = \exp\left(-\frac{||p - p'||^2}{2\sigma^2}\right), \tag{1}$$

where $|N(p)|$ is the cardinality of $N(p)$, and σ is a hyper-parameter of the RBF kernel. By increasing the point set size, the two set abstraction layers in the geometry extractor can learn local features at different contextual scales.

With the locally grouped point features, the feature extraction module then conducts global feature propagation to enrich these local representations. Due to irregular domain and disordering of the point cloud, the global propagation is realized via a permutation invariant attention layer, which is widely used in neural language processing (Vaswani et al., 2017; Yang et al., 2019). The attention layer first computes an attention map A, which records the pair-wise relationships between points (Guo et al., 2021), and the global propagation is performed by enriching any given point representation with the information of other points under the guidance of the attention map. As depicted in Fig. 2(c), given a point feature F_p, which is derived by applying linear transformation to F_p^{loc}, the information contained in other points is delivered to F_p with the relation weights contained in the attention map. In this case, a point representation is strengthened by the information in semantically and geometrically related features. Eq. (2) defines the procedure of obtaining a feature F_p^{global} that is enriched with global feature propagation.

$$F_p^{\text{global}} = \sum_{p' \in N(p)} A\left(p, p'\right) F_{p'}, \tag{2}$$

where $F_{p'}$ denotes the feature vector belonging to point p'.

The feature extraction module can now merge F^{loc} and F^{global} into a more representative geometric feature by using the following equation:

$$F_p^{\text{geo}} = F_p^{\text{loc}} + \alpha F_p^{\text{global}}. \tag{3}$$

In Eq. (3), instead of adopting direct summation as many previous works (Vaswani et al., 2017; Yang et al., 2019), we set a learnable factor α to control the contribution of F^{global}, since our geometry extractor should support various parsing tasks that may have different dependencies on global geometric context.

Summarily, built on PointNet++ (Qi et al., 2017), the geometry extraction module can not only learn features with multi-scale context via the set abstractions layers and MSG strategy like PointNet++, but also can enrich every point feature with information transferred across the entire point cloud by the global propagation step. The geometry extraction module therefore learns and enriches point features by propagating feature information at both local and global levels, and further fuses the global context information with local clues with a learnable parameter. The features learnt with local and global context information can benefit the sub-tasks of indoor semantic reconstruction to different degrees. For example, predicting the camera pose may require more global layout structure information, while estimating the pose of an object may need more local shape clues.

In our geometry extractor, two feature extraction modules are first adopted to learn point-wise geometric features, and a set abstraction layer with global average pooling merges the point-wise features into a representative geometric feature vector that encodes both local shape and global context for sub-tasks. With the geometry extractor as the backbone, the layout estimator and object detector of the proposed

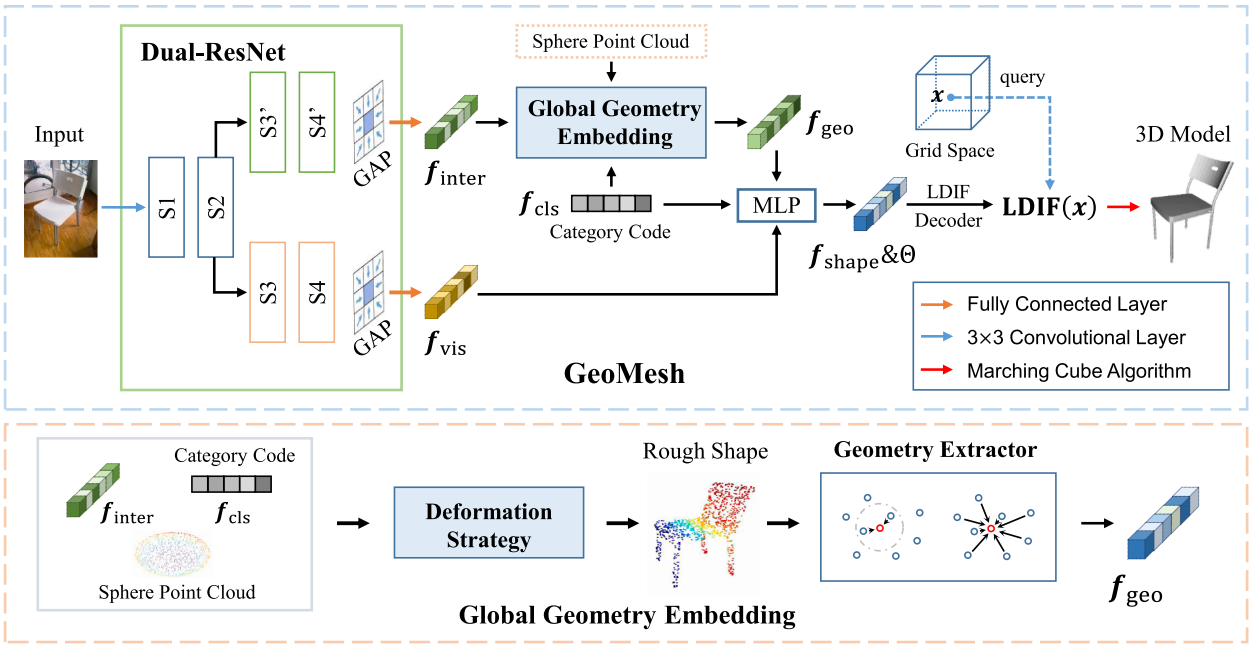

Fig. 4. The Architecture of GeoMesh. GAP refers to the global average pooling. Symbols of S* denote the stages of ResNet18, where S3, S4 are for f_{vis} and S3', S4' for f_{geo}. The sphere point cloud is a pre-defined 2562-point template, and it is deformed via the deformation strategy to generate a point cloud that fits the rough object shape. The rough shape point cloud is then fed into the geometry extractor to yield the global geometry feature f_{geo}.

GeoRec method are constructed as follows.

3.3. Geometry-enhanced layout estimation and object detection

In GeoRec, the geometry-enhanced layout estimator (GeoLE) and geometry-enhanced object detector (GeoOD) are built with two independent geometry extractor backbones and several multi-layer perceptrons (MLPs) to parse 3D geometric properties, including the 3D layout cuboid, camera pose, 3D object bounding boxes and object poses. Fig. 3 offers details regarding the utilization of MLPs by GeoLE and GeoOD to decode the geometric feature provided by the geometry extractor, and output trainable parameters that determine the layout properties and the object attributes.

In line with prior works (Nie et al., 2020; Zhang et al., 2021), GeoOD decodes the geometry feature with the help of the category code offered by 2D detection results, and the 3D geometric attributes, such as the layout and the 3D object bounding box, are parameterized into learnable targets with a world coordinate system located at camera center. The vertical axis of the world system is set perpendicular to the floor, and the forward axis is directed toward the camera. In the world system, the camera pose can be represented with corresponding pitch (β) and roll (γ) angles by Eq. (4) as shown below.

$$\boldsymbol{R}(\beta,\gamma)=\begin{bmatrix}\cos(\beta) & -\cos(\gamma)\sin(\beta) & \sin(\beta)\sin(\gamma)\\ \sin(\beta) & \cos(\beta)\cos(\gamma) & -\cos(\beta)\sin(\gamma)\\ 0 & \sin(\gamma) & \cos(\gamma)\end{bmatrix}. \tag{4}$$

The 3D layout cuboid can be decomposed into the center $\boldsymbol{c}^{\text{lay}}\in\mathbb{R}^3$, the spatial size $\boldsymbol{s}^{\text{lay}}\in\mathbb{R}^3$, and the orientation angle $\theta^{\text{lay}}\in[-\pi,\pi)$. With the object pose as the orientation angle θ^{obj}, a 3D bounding object box can be factorized similarly, whereas the 3D center $\boldsymbol{c}^{\text{obj}}_{\text{3d}}$ is differently determined by the corresponding 2D projection $\boldsymbol{c}^{\text{obj}}_{\text{2d}}\in\mathbb{R}^2$ and the distance $d\in\mathbb{R}$ to the camera center, in which case the 3D center of an object bounding box can be restored by Eq. (5).

$$\boldsymbol{c}^{\text{obj}}_{\text{3d}}=\boldsymbol{R}^{-1}(\beta,\gamma)\cdot d\cdot\frac{\boldsymbol{K}^{-1}[\boldsymbol{c}^{\text{obj}}_{\text{2d}},1]^T}{\left\|\boldsymbol{K}^{-1}[\boldsymbol{c}^{\text{obj}}_{\text{2d}},1]^T\right\|_2}, \tag{5}$$

where K is the camera intrinsic matrix. As noted by Huang et al. (2018a), Eq. (5) makes it possible to force a 2D-3D consistency during training to help stabilize the 3D object bounding box estimation. With the 2D object detection results, the estimation stabilization can be further strengthened by learning an offset $\boldsymbol{\delta}$ from the projection center $\boldsymbol{c}^{\text{obj}}_{\text{2d}}$ to the detected 2D box center. As $\boldsymbol{R}(\beta,\gamma)$ is estimated by the layout estimator, the estimation improvement brought by the geometry-enhanced layout estimator can also assist in better 3D object detection. The effect of the layout estimator on GeoOD is validated and discussed in Section 4.5.

As shown in Fig. 3, GeoLE decides $(\beta,\gamma,\boldsymbol{c}^{\text{lay}},\boldsymbol{s}^{\text{lay}},\theta^{\text{lay}})$ for layout cuboid and camera pose, while GeoOD estimates $(\boldsymbol{\delta},d,\boldsymbol{s}^{\text{obj}},\theta^{\text{obj}})$ for 3D object bounding boxes and object poses. The objective function for supervising the training of the GeoLE given a scene with N objects can be formulated as

$$\mathcal{L}_{\text{GeoLE}}=\mathcal{L}_{\beta}+\mathcal{L}_{\gamma}+\mathcal{L}_{\boldsymbol{c}^{\text{lay}}}+\mathcal{L}_{\boldsymbol{s}^{\text{lay}}}+\mathcal{L}_{\theta^{\text{lay}}}, \tag{6}$$

while the loss function for GeoOD is defined as

$$\mathcal{L}_{\text{GeoOD}}=\frac{1}{N}\sum_{n=1}^{N}\left(\mathcal{L}_{\boldsymbol{\delta}_n}+\mathcal{L}_{d_n}+\mathcal{L}_{\boldsymbol{s}^{\text{obj}}_n}+\mathcal{L}_{\theta^{\text{obj}}_n}\right). \tag{7}$$

Drawing on the spirits of previous works, we use a hybrid loss that combines classification and offset regression for stably learning these target parameters. Details of the loss function can be referred to object detection works (Ren et al., 2015; Mousavian et al., 2017; Huang et al., 2018a; Qi et al., 2018).

3.4. Geometry-enhanced mesh generation

Object instance modeling is a challenging task in indoor semantic reconstruction, because indoor objects are often accompanied by issues of partial occlusion and truncation. For high-quality and efficient object reconstruction, the deep implicit function (DIF) representation was adopted in recent works to reconstruct objects. The DIF, which is formulated as $f(\boldsymbol{x},\boldsymbol{z})$, is an implicit 3D representation that decides whether a query 3D location $\boldsymbol{x}$ lies inside or outside the object surface with the clues offered by the latent shape code $\boldsymbol{z}$. Compared with the

direct learning of voxel grid or topological modification on a template, the DIF representation can provide more high-quality detail recovery with flexible efficiency. Nevertheless, the DIF used in existing methods can still encounter difficulties in the reconstruction of highly occluded objects, due to the lack of global object structure clues in the latent shape code z. To counter the impact of occlusion, the proposed geometry-enhanced mesh generator (GeoMesh) learns LDIF$(\boldsymbol{x}, z, \boldsymbol{\Theta})$ (a kind of DIF representation defined by Genova et al. (2020)) with global geometry information embedded into the latent shape embedding z and the LDIF analytic code $\boldsymbol{\Theta}$. Object meshes can be obtained by classifying the points on a regular grid into inside/outside category with LDIF and running the marching cube algorithm (Lorensen and Cline, 1987). The same with Total3D and Im3D, we adopted the challenging SUN RGB-D dataset (Song et al., 2015)for single-view indoor semantic reconstruction. As the indoor scene data in SUN RGB-D dataset have no corresponding mesh models, we follow Total3D and Im3D to train the GeoMesh with the object-centric RGB images provided by the Pix3D dataset (Sun et al., 2018)for single object reconstruction, and apply the trained GeoMesh to the RGB data in SUN RGB-D dataset.

The GeoMesh extracts implicit shape embedding z from object-centric color images and the related category codes, which come from the 2D object detection results. As portrayed in Fig. 4, GeoMesh first derives a visual appearance descriptor $\boldsymbol{f}_{\text{vis}}$ and a global geometry feature $\boldsymbol{f}_{\text{geo}}$ from the input image, and yields the geometry-enhanced implicit shape embedding $\boldsymbol{f}_{\text{shape}}$ and the analytic code $\boldsymbol{\Theta}$ by fusing $\boldsymbol{f}_{\text{vis}}$ and $\boldsymbol{f}_{\text{geo}}$ with the category code $\boldsymbol{f}_{\text{cls}}$. The shape embedding $\boldsymbol{f}_{\text{shape}}$ and the analytic code $\boldsymbol{\Theta}$ are then fed into the LDIF decoder to generate the LDIF representation of an object shape.

At the shape encoding stage, we modified the ResNet18 (He et al., 2016) into a two-branch backbone (namely Dual-ResNet) to produce $\boldsymbol{f}_{\text{vis}}$ and an intermediate feature $\boldsymbol{f}_{\text{inter}}$, which is used for inferring $\boldsymbol{f}_{\text{geo}}$. The standard ResNet18 is composed of four encoding stages and a global average pooling (GAP) followed by a fully connected convolution layer (FC). The Dual-ResNet inherits the first two encoding stages for low-level feature learning, and duplicates the remaining structure of the ResNet18 into two branches for the generation of $\boldsymbol{f}_{\text{vis}}$ and $\boldsymbol{f}_{\text{inter}}$. The global geometry embedding step then deforms a sphere point cloud fit the point cloud to the rough 3D shape of the target object with $\boldsymbol{f}_{\text{inter}}$ and $\boldsymbol{f}_{\text{cls}}$, and infers the global geometry features $\boldsymbol{f}_{\text{geo}}$ from the rough shape with the geometry extractor defined in Section 3.2. In detail, the deformation strategy is deployed as Eq. (8), and it iterates twice to shift each point $\boldsymbol{p}$ with the offset derived by a multi-layer perceptron (MLP) from the point coordinate and $\boldsymbol{f}_{\text{inter}}$.

$$\boldsymbol{p}_{\text{deform}} = \boldsymbol{p} + \text{MLP}([\boldsymbol{p}, \boldsymbol{f}_{\text{inter}}, \boldsymbol{f}_{\text{cls}}]). \tag{8}$$

Given the deformed point cloud as P' and the ground-truth point set P, the deformation is learned with the supervision of Chamfer distance loss (Mescheder et al., 2019) defined as follows.

$$\mathscr{L}_{\text{chamfer}} = \frac{1}{|P|}\sum_{\boldsymbol{p}\in P}\min_{\boldsymbol{p}'\in P'}\|\boldsymbol{p}-\boldsymbol{p}'\|_2^2 + \frac{1}{|P'|}\sum_{\boldsymbol{p}'\in P'}\min_{\boldsymbol{p}\in P}\|\boldsymbol{p}'-\boldsymbol{p}\|_2^2. \tag{9}$$

Subsequently, the geometry extractor encodes the global geometry structure described by the deformed point cloud into $\boldsymbol{f}_{\text{geo}}$. With $\boldsymbol{f}_{\text{vis}}$ that contains appearance details and $\boldsymbol{f}_{\text{geo}}$ that encodes global geometry, the geometry-enhanced shape embedding $\boldsymbol{f}_{\text{shape}}$ and the LDIF analytic code $\boldsymbol{\Theta}$ are calculated by merging $\boldsymbol{f}_{\text{vis}}, \boldsymbol{f}_{\text{geo}}$ and $\boldsymbol{f}_{\text{cls}}$ with concatenation and multi-layer perceptrons.

The LDIF decoder (Genova et al., 2020) estimates LDIF$(\boldsymbol{x}, z, \boldsymbol{\Theta})$ with the shape and analytic codes ($\boldsymbol{f}_{\text{shape}}$ and $\boldsymbol{\Theta}$) that are enhanced with global geometry information as the input. The learning of LDIF$(\boldsymbol{x}, \boldsymbol{f}_{\text{shape}}, \boldsymbol{\Theta})$ is supervised by $\mathscr{L}_{\text{LDIF}}$ (Eq. (10)), which measures the accuracy of LDIF$(\boldsymbol{x}, \boldsymbol{f}_{\text{shape}}, \boldsymbol{\Theta})$ in deciding the inside/outside of a ground-truth 3D shape.

$$\begin{aligned}\mathscr{L}_{\text{LDIF}} &= \lambda_{\text{us}}\mathscr{L}_{\text{us}} + \lambda_{\text{ns}}\mathscr{L}_{\text{ns}} \quad \text{and} \quad \mathscr{L}_{\text{us/ns}} \\ &= \frac{1}{|S_{\text{us/ns}}|}\sum_{\boldsymbol{p}\in S_{\text{us/ns}}}\left\|\text{sigmoid}\left(\tau\text{LDIF}(\boldsymbol{p}, \boldsymbol{f}_{\text{shape}}, \boldsymbol{\Theta})\right) - \text{GT}(\boldsymbol{p})\right\|_2^2,\end{aligned} \tag{10}$$

$$\text{where} \quad \text{sigmoid}(x) = \frac{1}{1+e^{-x}} \quad \text{and} \quad \text{GT}(\boldsymbol{p}) = \begin{cases} 0, & \boldsymbol{p} \text{ is inside} \\ 1, & \boldsymbol{p} \text{ is outside.} \end{cases}$$

In Eq. (10), $|\cdot|$ and GT refer to the cardinality of a set and a label indicator function, respectively, and $\mathscr{L}_{\text{LDIF}}$ is a weighted combination of a uniform sample point loss $\mathscr{L}_{\text{us}}$ and a near surface sample point loss $\mathscr{L}_{\text{ns}}$. $\mathscr{L}_{\text{us}}$ refines the prediction of points (set S_{us}) that are uniformly sampled within the bounding box of the ground-truth shape, while $\mathscr{L}_{\text{ns}}$ focuses on near-surface points (set S_{ns}) that are selected by the sampling strategy used by Genova et al. (2019). The LDIF value is scaled by the hyper-parameter τ, and a sigmoid activation function is applied to compress the scaled LDIF into $(0,1)$. With 1 and 0 respectively denoting "outside" and "inside" labels, respectively, the compressed value can represent the probability that a point falls outside.

The objective function that trains the GeoMesh model is formulated as follows.

$$\mathscr{L}_{\text{GeoMesh}} = \lambda_{\text{chamfer}}\mathscr{L}_{\text{chamfer}} + \lambda_{\text{us}}\mathscr{L}_{\text{us}} + \lambda_{\text{ns}}\mathscr{L}_{\text{ns}}, \tag{11}$$

where the weights λ_{chamfer} and λ_{us} are set to 1, while λ_{ns} is 0.4 in practice.

4. Experiments and analysis

We conducted experiments on two commonly-used benchmark datasets, Pix3D (Sun et al., 2018) (a large-scale object reconstruction dataset) and SUN RGB-D (Song et al., 2015) (a challenging benchmark for indoor scene understanding). The 3D reconstruction quality of GeoMesh was validated on the Pix3D, and the accuracy of the layout and camera pose estimated by GeoLE and the 3D object detection performance of GeoOD were evaluated on the SUN RGB-D dataset. For the holistic indoor semantic reconstruction, GeoMesh trained on the Pix3D was directly transferred to the SUN RGB-D dataset to provide restored indoor object meshes, which will be arranged in the room space estimated by GeoLE with the object geometric attributes parsed by GeoOD.

In the following, we first elaborate the datasets and experimental settings, and present the performance comparison on different datasets. We then provide a qualitative analysis regarding indoor semantic reconstruction. Finally, the ablation studies are presented to investigate every sub-task component of the GeoRec for comprehensive understanding.

4.1. Datasets and implementation protocol

4.1.1. Dataset and metrics

Pix3D: The Pix3D dataset (Sun et al., 2018) is a large-scale benchmark for shape-related tasks, *e.g.*,3D object mesh reconstruction. This dataset contains 10,069 images paired with 395 precisely aligned furniture models, which are classified into 9 categories. The train/test split setting was kept in line with previous works (Gkioxari and Malik, 2019; Nie et al., 2020). The data pre-processing pipeline in Im3D (Zhang et al., 2021) was adopted to get watertight meshes for GeoMesh training, and the GeoMesh was validated with the original data.

SUN RGB-D: The SUN RGB-D (Song et al., 2015) is a challenging dataset for holistic indoor scene understanding, with 10335 indoor scene RGB-D images captured by diverse sensors. Each image in the SUN RGB-D dataset is attached with pixel-level 2D semantic segmentation annotation, 3D room layout, 2D and 3D bounding boxes, as well as object orientations. The SUN RGB-D offers an official split setting that divides the dataset into a training set that includes 5285 RGB-D images and a testing set that contains the other 5050 images. For indoor scene semantic reconstruction, we used the label mapping defined by Total3D

Table 1
Quantitative comparison for object mesh reconstruction on Pix3D dataset. The evaluation metric is the Chamfer distance (lower is better), which is computed with 10 K points sampled from the 3D object mesh predictions that are aligned with the ground-truth via ICP algorithm. MGN and LIEN are the reconstruction generators used in Total3D and Im3D, respectively. The best results are marked in bold. (unit:10^{-3}).

Category	Data	bed	bookcase	chair	desk	sofa	table	tool	wardrobe	mean
AtlasNet (Groueix et al., 2018)	RGB	9.03	6.91	8.37	8.59	6.24	19.46	6.95	4.78	8.79
TMN (Pan et al., 2019)	RGB	7.78	5.93	6.86	7.08	4.25	17.42	4.13	4.09	7.19
MGN (Nie et al., 2020)	RGB	5.62	7.27	7.66	6.81	6.92	12.72	3.51	4.95	6.93
LIEN (Zhang et al., 2021)	RGB	5.38	5.07	5.66	9.89	3.37	13.85	3.57	3.06	6.23
GeoMesh (Ours)	RGB	**4.00**	**4.19**	**5.35**	**8.93**	**3.28**	**10.40**	**2.88**	**2.48**	**5.19**

(Nie et al., 2020), which maps the SUN RGB-D object categories to the Pix3D furniture classes, to transfer the GeoMesh trained on Pix3D to SUN RGB-D.

Metrics: We evaluated the performance of the every sub-task module in GeoRec with the standard metrics used in previous works, such as Total3D (Nie et al., 2020) and Im3D (Zhang et al., 2021). Specifically, the object reconstruction quality is measured by the Chamfer distance; 3D object detection by the average precision (AP) with box IoU threshold set as 0.15; the layout estimation accuracy by the average 3D intersection over union (IoU); and the camera pose by the mean absolute error.

4.1.2. Implementation settings

The GeoRec method was implemented under Pytorch framework (Paszke et al., 2019) with each sub-task module, i.e., GeoLE, GoeOD and GeoMesh, individually trained on a NVIDIA TITAN RTX GPU. Following Im3D, we used the predictions provided by the 2D detector of Total3D as the 2D object detection input for GeoOD and GeoMesh. Each sub-task module in the GeoRec was optimized by Adam optimizer with initial learning rate 2×10^{-4} and weight decay 1×10^{-4}. The learning rate was scaled by 0.5 when the loss stopped decreasing for a given period, which is 30 epochs for GeoLE and GeoOD, while 50 for GeoMesh. The overall training period lasted 300 epochs for training GeoLE and GeoOD on SUN RGB-D dataset, and GeoMesh on Pix3D. The point cloud generated by the depth image is uniformly sub-sampled into 2500 and 5000 points to serve as the input of GeoLE and GeoOD, respectively. To apply GeoMesh to the SUN RGB-D, the RGB images were cropped by the 2D object bounding boxes provided by Total3D to generate object-centric input for GeoMesh, and the cropped images were resized into 256×256, which is the image size used for training GeoMesh with Pix3D. Color jitter was used for augmenting data when training GeoMesh, while random horizontal flipping was set during the training of every sub-task module. For the recovery of the indoor scene depicted by a given RGB-D image in SUN RGB-D dataset, all the sub-task modules of GeoRec were implemented jointly to simultaneously infer object meshes and the 3D geometric attributes of the layout and objects, and the semantic 3D model of the indoor scene was built by arranging object models in the estimated layout space with alignment to the object geometric properties.

Fig. 5. Visualization results of object shapes reconstructed by AltasNet, MGN, LIEN and the proposed GeoMesh on Pix3D dataset. It can be seen that GeoMesh preserves more geometric details and achieves higher reconstruction completeness with stronger robustness to occlusion.

Table 2
Quantitative comparison of the estimation of 3D room layout and camera pose. The best results are marked in bold. (↑: higher is better, ↓: lower is better).

Method	Data	3D Layout IoU↑ (unit: %)	Cam pitch↓	Cam roll↓
3DGP (Choi et al., 2013)	RGB	19.2	-	-
Hedau (Hedau et al., 2009)	RGB	-	33.85	3.45
HoPR (Huang et al., 2018b)	RGB	54.9	7.60	3.12
CooP (Huang et al., 2018a)	RGB	56.9	3.28	2.19
Total3D (Nie et al., 2020)	RGB	59.2	3.15	2.09
Im3D (Zhang et al., 2021)	RGB	64.4	2.98	2.11
ImVoxelNet (Rukhovich et al., 2021)	RGB	59.3	2.63	1.96
GeoRec(Ours)	Depth	**77.1**	**1.80**	**1.71**

4.2. Mesh reconstruction results on Pix3D

To investigate the performance of single object reconstruction with an object-centric color image, the mesh prediction module (GeoMesh) of the GeoRec, was evaluated on the challenging large-scale 3D shape recovery benchmark, i.e., the Pix3D dataset. The performance of GeoMesh is compared with several state-of-the-art mesh reconstruction methods, including AtlasNet (Groueix et al., 2018), TMN (Pan et al., 2019), MGN proposed in Total3D (Nie et al., 2020), and LIEN (Zhang et al., 2021) used in Im3D, and the quantitative comparison results are reported in Table 1, where the input of all the compared methods includes a cropped object-centric image and a one-hot category code.

As shown in Table 1, the proposed GeoMesh achieves the lowest mean chamfer distance value (5.19×10^{-3}) with high 3D shape reconstruction accuracy, outperforming the other methods across object categories. Among the methods listed in Table 1, LIEN and GeoMesh are two methods that learn LDIF representation for 3D geometry recovery. However, our GeoMesh outperforms LIEN by a large margin in terms of all categories and the mean chamfer distance, demonstrating that the global geometry structure information encoded in GeoMesh is effective for high-quality object geometry restoration. To better reveal the quality of the 3D object shape reconstructed by different methods in terms of completeness and fineness, some qualitative comparison results are displayed in Fig. 5.

In Fig. 5, we compared the visualization results of GeoMesh with those of AltasNet, MGN and LIEN for qualitative analysis. Fig. 5 shows that GeoMesh is capable of object geometry restoration with high completeness and strong robustness to occlusion, whereas LIEN is sensitive to occlusion. The topology deformation methods, including AtlasNet and MGN, often face deformation artifacts (*e.g.,* angular object surfaces), because of the numerous degrees of freedom for mesh vertexes. Specifically, comparing the reconstructed chair and sofa models of the listed methods in the first two columns, it can be observed that, although all methods successfully models the objects depicted in the images, the proposed GeoMesh captures the shapes of the chair and sofa with smoother object surfaces and the reconstructed geometric structures of GeoMesh are also better fitted to the physical objects. The object mesh models in other columns validate that GeoMesh can still guarantee reconstruction completeness when the objects in the input images are partially occluded. For example, when the tables are blocked by chairs as shown in the third and fourth columns, GeoMesh clearly delineates the complete 3D geometry of the tables, while LIEN yields incomplete or wrong meshes due to the interference of the chairs. Similar phenomenon can also be found in the fifth column where the reconstruction of bookcase is challenged by piles of books. The reconstruction quality of the double-deck table in the last column further indicates the efficacy of GeoMesh for modeling object with holes. In contrast to our GeoMesh, other methods encounters difficulties more or less in recovering the accurate table structure due to the influence of the hole or the occlusion. From Fig. 5, apart from geometry completeness and robustness to occlusion, GeoMesh also provides more straight lines and smoother curves, which are the general characteristics of man-made objects. Therefore, benefiting from the effective learning of global geometry information, our GeoMesh simultaneously achieves occlusion-robust reconstruction and geometry detail preservation.

4.3. Results on SUN RGB-D

4.3.1. 3D geometric property parsing

In this section, we compare and discuss the performance of our GeoRec and other existing indoor understanding methods for 3D geometric property parsing with SUN RGB-D dataset.

3D layout estimation: Table 2 presents the 3D layout IoU scores of several methods, and the prediction error of the camera pitch and roll angles. With respect to layout estimation, our GeoRec method exceeds other methods by at least 12.7% in terms of 3D IoU metric. Our GeoRec also obtains the most accurate inference of camera pose with the lowest error values. The outstanding performance in the estimation of the layout and camera pose can be attributed to the effective exploitation of the geometry information provided in depth data. The global propagation in the geometry extractor also makes our GeoRec more aware of the holistic scene structure, which is important for capturing the layout and camera pose. More detailed discussion on the influence of the global propagation can be found in Section 4.4.

3D object detection: Same with previous works, the 3D object detection performance is evaluated via the mean average precision (mAP) metric. Table 3 displays the numeric comparison to the existing 3D object detectors.

As shown in Table 3, our GeoRec significantly improves the 3D object detection accuracy and achieves the highest mAP score, which surpasses those of other methods by at least 25.43%. GeoRec also consistently outperforms other methods by a remarkable margin on all categories.

Table 3 also reveals that GeoRec is more robust to object scale variation than other methods with the accurate geometric clues contained in depth data, since GeoRec accurately detects not only large objects (*e.g.,* the bed and sofa), but also small items, such as the lamp. For instance, when other methods detects the lamp with AP scores that are generally lower than 20%, due to the localization hardness caused by

Table 3
Category-wise and mean AP scores for 3D object detection on the SUN RGB-D dataset (the higher is better). CooP* is the model retrained with all categories by the authors of Total3D for a fair comparison. The best results are marked as bold. Dresser and nightstand are abbreviated as Drsr and Ntsd. (unit: %).

Method	Data	Bed	Chair	Sofa	Table	Desk	Drsr	Ntsd	Sink	Cabinet	Lamp	mAP
3DGP (Choi et al., 2013)	RGB	5.62	2.31	3.24	1.23	-	-	-	-	-	-	-
HoPR (Huang et al., 2018b)	RGB	58.29	13.56	28.37	12.12	4.79	13.71	8.80	2.18	0.48	2.41	14.47
Coop (Huang et al., 2018a)	RGB	63.58	17.12	41.22	26.21	9.55	4.28	6.34	5.34	2.63	1.75	17.80
CooP* (Huang et al., 2018a)	RGB	57.71	15.21	36.67	31.16	19.90	15.98	11.36	15.95	10.47	3.28	21.77
PerspectiveNet (Huang et al., 2019)	RGB	79.69	40.42	62.35	44.12	20.19	27.38	35.16	41.35	1.70	13.14	36.55
Total3D (Nie et al., 2020)	RGB	60.65	17.55	44.90	36.48	27.93	21.19	17.01	18.50	14.51	5.04	26.38
Im3D (Zhang et al., 2021)	RGB	89.34	35.14	69.10	57.37	49.03	29.27	41.34	33.81	33.93	11.90	45.02
GeoRec(Ours)	RGBD	**95.50**	**48.67**	**85.23**	**69.88**	**63.87**	**58.01**	**75.78**	**76.42**	**79.53**	**51.59**	**70.45**

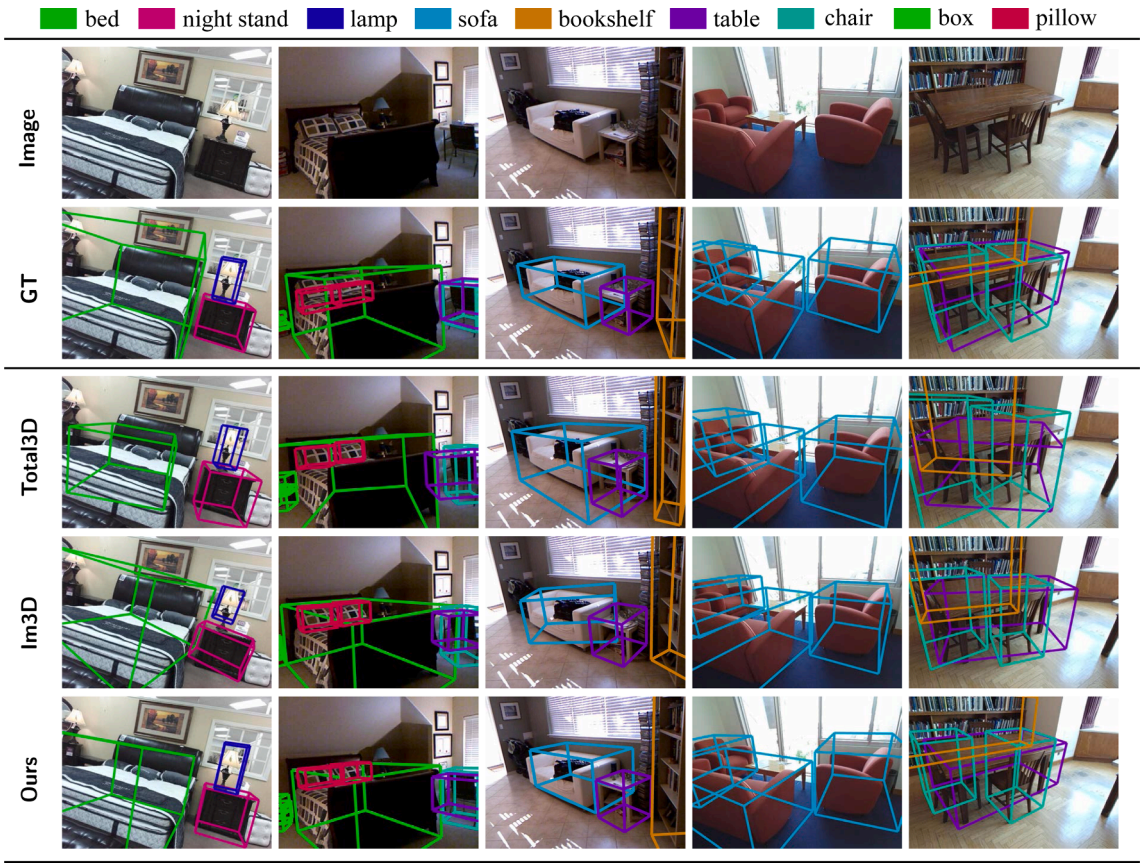

Fig. 6. Qualitative comparison between Total3D (Nie et al., 2020), Im3D (Zhang et al., 2021) and our GeoRec on SUN RGB-D dataset in terms of 3D o.bject detection.

Table 4
Category-wise and mean AP scores computed with 0.25 IoU threshold for 3D object detection on the SUN RGB-D dataset. Except VoteNet that only used depth data, all the other 3D object detection methods in the table utilized the RGB-D data. Bath, Bkshf, Drsr and Ntsd denote the bathtub, bookshelf, dresser and nightstand, respectively. The best results are marked in **bold** and the second best are underlined (unit: %).

Method	Data	Bath	Bed	Bkshf	Chair	Desk	Drsr	Ntsd	Sofa	Table	Toilet	mAP
DSS (Song and Xiao, 2016)	RGBD	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
2d-driven (Lahoud and Ghanem, 2017)	RGBD	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
COG (Ren and Sudderth, 2016)	RGBD	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	51.3	70.1	47.6
COG-Latent (Ren and Sudderth, 2018)	RGBD	76.2	73.2	32.9	60.5	34.5	13.5	30.4	60.4	**55.4**	73.7	51.0
PointFusion (Xu et al., 2018)	RGBD	37.3	68.6	37.7	55.1	17.2	24.0	32.2	53.8	31.0	83.8	44.1
F-PointNet (Qi et al., 2018)	RGBD	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	90.9	54.0
VoteNet (Qi et al., 2019)	Depth	74.4	83.0	28.8	75.3	22.0	29.8	**62.2**	64.0	47.3	90.1	57.7
EPN (Ahmed and Chew, 2020)	RGBD	79.4	**88.2**	32.1	17.0	37.4	**53.7**	50.0	65.3	53.3	**95.8**	57.2
MBDF-Net (Tan et al., 2021)	RGBD	**81.5**	84.7	33.0	**77.3**	31.2	29.0	57.7	65.6	49.9	85.5	**59.5**
Ours	RGBD	62.4	86.1	**49.7**	38.0	40.5	45.5	59.3	**69.9**	52.3	79.3	58.3

the slim shape and the influence of other adjacent larger objects (*e.g.,* the bed and the nightstand), our GeoRec produces an AP score of 51.59% for the lamp category, which demonstrates the capability of GeoRec for locating the actual 3D positions of small objects. For perceptual inspection of the detection performance of GeoRec, we presents visualization results in Fig. 6.

Considering current multi-task indoor scene parsing methods listed in Table 3 perform 3D object detection with only RGB images, while our method utilizes RGB-D data, we therefore add comparison between our method and other RGB-D approaches that focus on 3D object detection in Table 4. It can be seen that the GeoOD module of our GeoRec achieves competitive performance with respect to existing single-task 3D object detectors. It should be noted that the GeoRec is a multi-task system that is designed for effectively achieving several sub-tasks for indoor semantic reconstruction, while these RGB-D methods are task-specifically proposed for 3D object detection. Therefore, the GeoOD may not be as complex and powerful as current 3D RGB-D detectors, but it is effective under the multi-task framework for indoor semantic reconstruction. For a user with sufficient computational resources, the GeoOD can be replaced with higher-performance RGB-D detectors.

Fig. 6 compares the visualization results of GeoRec with those of Total3D and Im3D, two state-of-the-art indoor semantic reconstruction methods. It can be seen that GeoRec produces 3D object bounding boxes that more compactly cover target objects with more accurate orientation than the other two methods. Compared to Total3D and Im3D that yield bounding boxes with either wrong orientation or improper size when detecting the beds and sofas in first four columns, the GeoRec encapsulates the beds and sofas with 3D boxes that more properly match the size and shape of the targets. Furthermore, the detection results in Fig. 6 reveal that the GeoRec is more robust to occlusion and truncation than

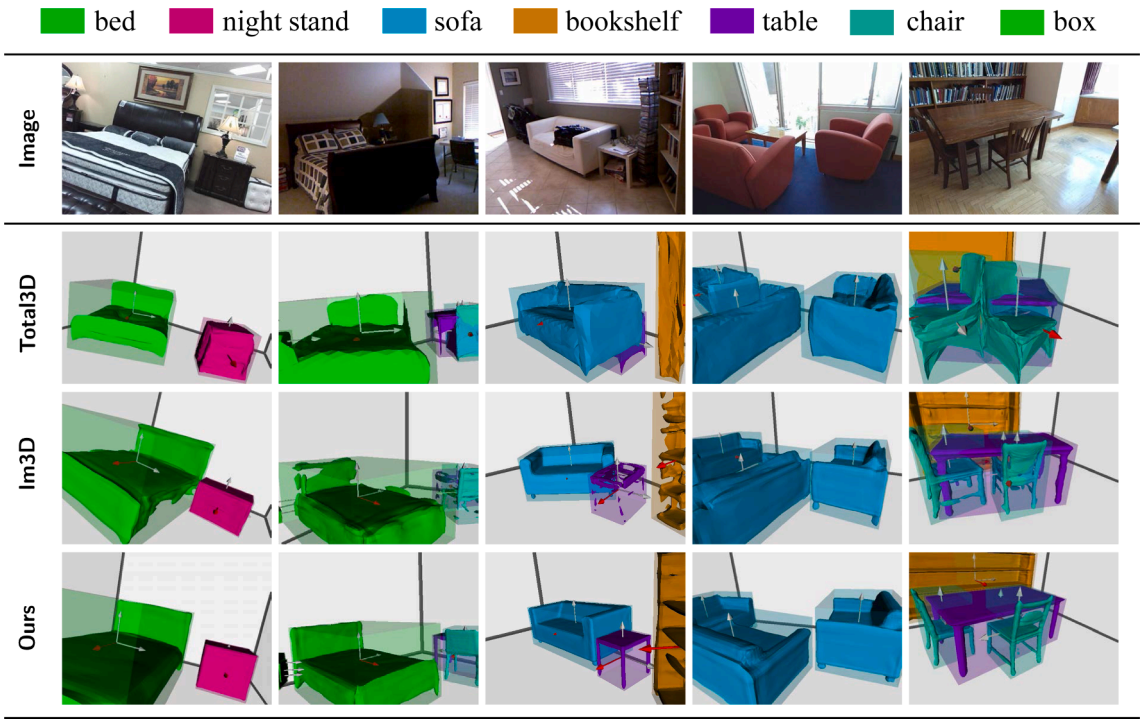

Fig. 7. Qualitative comparison between Total3D (Nie et al., 2020), Im3D (Zhang et al., 2021) and our GeoRec on SUN RGB-D dataset in terms of the 3D semantic reconstruction of indoor scenes. The 3D coordinate arrows on each object denote the estimated object orientation.

Table 5
Ablation study results for GeoMesh on Pix3D dataset. Geo refers to the global geometry embedding module, and GP means the global feature propagation used in the geometry extractor, which serves as a core part of the global geometry embedding module. The evaluation metric is the Chamfer distance computed with 10 K points sampled from the 3D object mesh predictions, which are aligned with the ground-truth via ICP algorithm. The lower the Chamfer distance values are, the better the reconstruction performance is. (unit:10^{-3}).

Method	Dual-ResNet	GeoEmb	GP	bed	bookcase	chair	desk	sofa	table	tool	wardrobe	mean
Baseline				5.38	5.07	5.66	9.89	3.37	13.85	3.57	3.06	6.23
GeoMesh		✓		5.04	5.79	5.90	9.39	3.54	14.15	3.22	2.50	6.19
GeoMesh		✓	✓	5.12	5.54	6.54	10.59	3.36	12.79	4.61	2.46	6.37
GeoMesh	✓	✓		4.73	3.67	5.61	9.80	3.29	11.63	2.79	2.39	5.49
GeoMesh	✓	✓	✓	4.00	4.19	5.35	8.93	3.28	10.40	2.88	2.48	5.19

Total3D and Im3D when locating the spatial positions of indoor objects. Taking the results in the last column for example, although the table (marked in purple) and chairs (marked in blue) are mutually occluded, and the lower part of the bookshelf (marked as yellow) is also blocked by the table, the GeoRec successfully captures these indoor objects with boxes that are nearly the same as human annotations, while Total3D and Im3D have trouble in correctly detecting the table with the localization confusion caused by the chairs.

4.3.2. Scene reconstruction

With the estimated 3D layout cuboid and detected 3D object bounding boxes, it is ready to perform scene reconstruction with SUN RGB-D data. As the SUN RGB-D dataset does not provide 3D models for indoor scenes, we directly applied the GeoMesh trained on Pix3D to SUN RGB-D data for object reconstruction without further fine-tuning. Qualitative comparison with existing indoor semantic methods is provided in Fig. 7. From Fig. 7, one can see that our method recovers indoor scenes with a more accurate estimation of the layout, camera pose, and 3D object bounding boxes. Especially, our GeoMesh demonstrates a stronger generalization ability than the object reconstruction methods proposed in Total3D and Im3D.

4.4. Ablation Study

In this section, we provide and discuss the ablation studies of GeoMesh, GeoOD and GeoLE with experiments on the Pix3D and SUN RGB-D datasets, for a deep insight into every sub-task component of the proposed GeoRec method.

4.4.1. 3D object reconstruction

To investigate the influence of the Dual-ResNet backbone and the global geometry embedding module in GeoMesh, we set the baseline as a model that adopts the ResNet18 as the backbone without the global geometry embedding module, in which case the baseline is the same with the LIEN used in Im3D (Zhang et al., 2021). Table 5 lists the results of models under different experiment settings.

Comparing the numeric results in the first three rows Table 5, we find that the global geometry embedding module (GeoEmb) without the Dual-ResNet can hardly reduce the reconstruction error. As Dual-ResNet works by providing two high-level feature vectors for global geometry embedding and visual appearance encoding, this phenomenon implies that learning confusion may be incurred if using the same high-level visual representation provided by the ResNet18 for the two different encoding targets. From the fourth row in Table 5, obvious performance improvement can be witnessed with the assistance of Dual-ResNet, which indicates the necessity of Dual-ResNet for exerting the global

Table 6
Ablation studies for GeoLE and GeoOD in terms of the estimation of 3D object bounding box, 3D room layout and camera pose. GP refers to the global feature propagation step in the geometry extractor, which is the backbone of GeoLE and GeoOD. (↑: higher is better, ↓: lower is better).

Setting		GeoOD	GeoLE		
RGB	GP	mAP↑	3D Layout IoU↑	Cam pitch↓	Cam roll↓
✓		67.96	74.83	2.41	2.14
✓	✓	67.99	75.06	2.41	2.13
		68.73	75.77	1.93	1.88
	✓	70.45	77.08	1.80	1.71

Table 7
The performance of the 3D object detector in the GeoRec with different layout estimators, where $LE_{Total3D}$ and LE_{Im3D} respectively refer to the layout estimators of Total3D and Im3D. Pitch err. and Pitch err. denote the estimation error of the camera pitch and roll angles. The object detection evaluation metric is the mAP scores. The best results are marked in **bold** and the second best are underlined(↓: lower is better).

	+GeoLE	+ $LE_{Total3D}$	+ LE_{Im3D}
Pitch err.↓	**1.80**	3.15	2.98
Roll err.↓	**1.71**	2.09	2.11
Obj Det. ↑	**70.45**	65.27	66.74

geometry embedding module. With the confusion issue alleviated by Dual-ResNet, the results in the last two row show that the global feature propagation (GP) step can further improve the power of global geometry embedding module. This can be attributed to the fact that the global propagation step is capable of delivering the shape information across the point features of the rough object point cloud, which enhance the final geometry embedding with rich global geometric structure information of the target object for improving the reconstruction quality. In this way, compared to the baseline the complete GeoMesh improves the reconstruction robustness and completeness with the cooperation of the Dual-ResNet and the global geometry embedding.

4.4.2. 3D geometric property parsing

For 3D geometric property parsing, we validated the necessity of the global propagation step for GeoLE and GeoOD, and we also studied the influence of extra color information offered by RGB images on the detection of the 3D object bounding box and the estimation of the layout cuboid and the camera pose. Corresponding numerical results are displayed in Table 6.

To study the influence of RGB information, the RGB values are taken as the input with the point coordinates for GeoLE and GeoOD. In Table 6, the results with and without RGB information reveal that additional RGB information generally helps little or even worsens the performance for 3D property parsing tasks. This may be because RGB information often has larger variation than geometry information due to the visual changes of indoor scenes, such as the illumination and object appearances. Therefore, the unstable color information can weaken the generalization of GeoLE and GeoOD for various room spaces, in which case the capability of the global feature propagation (GP) is also limited. When only using the geometric clues provided by depth data, both GeoOD and GeoLE achieves the best performance in Table 6 with global propagation step, which reveals that the global propagation step can strengthen the ability of GeoLE and GeoOD for fully excavating and leveraging the geometric information.

4.5. The joint effect of sub-modules for scene reconstruction

As GeoRec is a unified multi-task reconstruction system for single-view RGB- D images, and the sub modules need to not only accomplish different sub-tasks, but also provide essential information for other sub modules to jointly realize the reconstruction of a complete 3D indoor scene. In this section, we discuss the joint effect of sub-modules for indoor scene reconstruction.

The joint effect of GeoLE and GeoOD. As indicated by Eq. (5), the computation of a 3D box center predicted by GeoOD is closely related to the camera pose estimated by GeoLE. To validate the effect of layout estimator on the performance of the 3D object detector GeoOD, we conducted ablation studies by replacing GeoLE in our GeoRec with the layout estimators of Total3D (Nie et al., 2020) and Im3D (Zhang et al., 2021). The quantitative results are displayed in Table 7.

In Table 7, when replacing GeoLE with the layout estimators of Total3D (+ $LE_{Total3D}$) and Im3D (+ LE_{Im3D}), the camera pose estimation errors (pitch err. and roll err.) increase, while the accuracy of 3D object detection (Obj Det.) decreases. The results indicate that our GeoLE is closely related to the 3D object detector, and can improve the detection accuracy with more accurate camera estimation.

The joint effect of GeoOD and GeoMesh. In GeoRec, GeoOD provides the information of object position, size and pose for arranging the object meshes predicted by GeoMesh. Therefore, GeoOD and GeoMesh jointly determine the reconstruction quality of objects in the final 3D scene model. An example for illustrating this joint effect is given in Fig. 8. It can be seen that, although the shape of the sofa (marked in blue) is correctly modelled by Im3D and Total3D, the inaccurate object pose and size provided by the 3D object boxes still hampers the reconstruction quality of the sofa in the final 3D scene. Contrarily, with a more accurate 3D object box predicted by GeoOD, our GeoRec can recover the

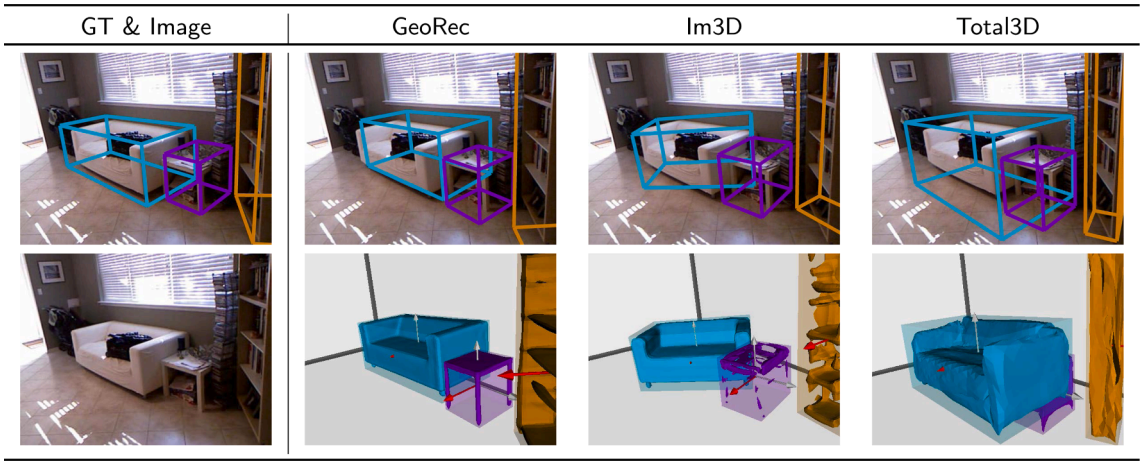

Fig. 8. The joint influence of the object detector and the mesh generator for high-quality object reconstruction in the 3D scene model..

sofa with higher quality in the scene model.

5. Conclusion

In this study, we presented a multi-task deep network (GeoRec) to learn geometry-enhanced feature representation from RGB-D data for robust indoor semantic 3D reconstruction. In GeoRec, we have presented a layout estimator (GeoLE) and an object detector (GeoOD) with a novel geometry extractor for accurate 3D geometric property parsing, and a new object mesh generator (GeoMesh) that captures occlusion-robust features for 3D mesh reconstruction. Extensive experiments on two large-scale challenging datasets demonstrate that GeoLE and GeoOD can effectively extract representative geometric features from depth data for estimating room layout and 3D object bounding boxes with high accuracy and strong robustness to indoor occlusion; while GeoMesh is capable of robust object reconstruction with richly encoded global geometry structure information. The experimental results also show that the three sub-task components of GeoRec can significantly exceed existing methods, and the semantic 3D models reconstructed by GeoRec are more accurate and complete than existing indoor semantic reconstruction methods.

In the future work, it would be promising to explore the combination of the semantic-assisted 3D indoor modeling and the general geometry-focused reconstruction, for more realistic recovery and holistic understanding of indoor scenarios. Although this work focuses on indoor semantic reconstruction, it is worth studying outdoor scene reconstruction in a similar semantic-assisted way, and replacing the depth data with high-quality LiDAR data may be beneficial for outdoor scene reconstruction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Acknowledgments This research was supported by the National Natural Science Foundation of China Project under Grant 42071370, the Open fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources and the Fundamental Research Funds for the Central Universities.

References

Ahmed, S.M., Chew, C.M., 2020. Density-based clustering for 3d object detection in point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10608–10617.

Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., Sivic, J., 2014. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3762–3769.

Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Nießner, M., 2019. Scan2cad: Learning cad model alignment in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2614–2623.

Avetisyan, A., Khanova, T., Choy, C., Dash, D., Dai, A., Nießner, M., 2020. SceneCAD: Predicting Object Alignments and Layouts in RGB-D Scans. pp. 596–612. https://doi.org/10.1007/978-3-030-58542-6_36.

Bansal, A., Russell, B., Gupta, A., 2016. Marr revisited: 2d-3d alignment via surface normal prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5965–5974.

Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S., 2013. Understanding indoor scenes using 3d geometric phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 33–40.

Dai, A., Nießner, M., 2018. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 452–468.

Dong, Z., Yang, B., Liu, Y., Liang, F., Li, B., Zang, Y., 2017. A novel binary shape context for 3d local surface description. ISPRS J. Photogramm. Remote Sens. 130, 431–452.

Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J., 2004. Recognizing objects in range data using regional point descriptors. In: Eurconference on Computer Vision. Springer, pp. 224–237.

Geiger, A., Wang, C., 2015. Joint 3d object and layout inference from a single rgb-d image. In: German Conference on Pattern Recognition. Springer, pp. 183–195.

Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T., 2020. Local deep implicit functions for 3d shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4857–4866.

Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W., Funkhouser, T., 2019. Learning shape templates with structured implicit functions. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7153–7163. https://doi.org/10.1109/ICCV.2019.00725.

Georgia Gkioxari, Jitendra Malik, J.J., 2019. Mesh r-cnn. ICCV 2019.

Gong, Z., Lin, H., Zhang, D., Luo, Z., Zelek, J., Chen, Y., Nurunnabi, A., Wang, C., Li, J., 2020. A frustum-based probabilistic framework for 3d object detection by fusion of lidar and camera data. ISPRS J. Photogram. Remote Sens. 159, 90–100.

Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M., 2018. A papier-mâché approach to learning 3d surface generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 216–224.

Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M., 2021. Pct: Point cloud transformer. Comput. Visual Media 7, 187–199.

Gupta, S., Arbeláez, P., Girshick, R., Malik, J., 2015. Aligning 3d models to rgb-d images of cluttered scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4731–4740.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hedau, V., Hoiem, D., Forsyth, D., 2009. Recovering the spatial layout of cluttered rooms, in: 2009 IEEE 12th international conference on computer vision, IEEE. pp. 1849–1856.

Hou, J., Dai, A., Nießner, M., 2020. Revealnet: Seeing behind objects in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2098–2107.

Huang, S., Chen, Y., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C., 2019. Perspectivenet: 3d object detection from a single rgb image via perspective points. arXiv preprint arXiv: 1912.07744.

Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N., Zhu, S.C., 2018a. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. arXiv preprint arXiv:1810.13049.

Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, S.C., 2018b. Holistic 3d scene parsing and reconstruction from a single rgb image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 187–203.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al., 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 559–568.

Izadinia, H., Shan, Q., Seitz, S.M., 2017. Im2cad. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5134–5143.

Johnson, A.E., Hebert, M., 1999. Using spin images for efficient object recognition in cluttered 3d scenes. IEEE Trans. Pattern Anal. Machine Intell. 21, 433–449.

Kang, Z., Yang, J., Yang, Z., Cheng, S., 2020. A review of techniques for 3d reconstruction of indoor environments. ISPRS Int. J. Geo-Informat. 9, 330.

Koppula, H.S., Anand, A., Joachims, T., Saxena, A., 2011. Semantic labeling of 3d point clouds for indoor scenes. In: Nips, p. 6.

Kulkarni, N., Misra, I., Tulsiani, S., Gupta, A., 2019. 3d-relnet: Joint object and relational network for 3d prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2212–2221.

Lahoud, J., Ghanem, B., 2017. 2d-driven 3d object detection in rgb-d images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4622–4630.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4558–4567.

Lee, C.Y., Badrinarayanan, V., Malisiewicz, T., Rabinovich, A., 2017. Roomnet: End-to-end room layout estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4865–4874.

Li, L., Khan, S., Barnes, N., 2019. Silhouette-assisted 3d object instance reconstruction from a cluttered scene. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0.

Li, L., Khan, S., Barnes, N., 2020a. Geometry to the rescue: 3d instance reconstruction from a cluttered scene. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 272–273.

Li, Y., Li, W., Tang, S., Darwish, W., Hu, Y., Chen, W., 2020b. Automatic indoor as-built building information models generation by using low-cost rgb-d sensors. Sensors 20, 293.

Lin, D., Fidler, S., Urtasun, R., 2013. Holistic scene understanding for 3d object detection with rgbd cameras. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1417–1424.

Lin, H., Wu, S., Chen, Y., Li, W., Luo, Z., Guo, Y., Wang, C., Li, J., 2021. Semantic segmentation of 3d indoor lidar point clouds through feature pyramid architecture search. ISPRS J. Photogramm. Remote Sens. 177, 279–290.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3d surface construction algorithm. ACM Siggraph Comput. Graphics 21, 163–169.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A., 2019. Occupancy networks: Learning 3d reconstruction in function space. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4455–4465. https://doi.org/10.1109/CVPR.2019.00459.

Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J., 2017. 3d bounding box estimation using deep learning and geometry. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 7074–7082.

Murez, Z., van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A., 2020. Atlas: End-to-end 3d scene reconstruction from posed images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, Springer. pp. 414–431.

Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J., 2020. 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.

Nie, Y., Hou, J., Han, X., Nießner, M., 2021. Rfd-net: Point scene understanding by semantic instance reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4608–4618.

Pan, J., Han, X., Chen, W., Tang, J., Jia, K., 2019. Deep mesh reconstruction from single rgb images via topology modification networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9964–9973.

Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S., 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Informat. Process. Syst. 8026–8037.

Qi, C.R., Chen, X., Litany, O., Guibas, L.J., 2020. Imvotenet: Boosting 3d object detection in point clouds with image votes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4404–4413.

Qi, C.R., Litany, O., He, K., Guibas, L.J., 2019. Deep hough voting for 3d object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9277–9286.

Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J., 2018. Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 918–927.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Informat. Process. Syst. 30.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Informat. Process. Syst. 28, 91–99.

Ren, Z., Sudderth, E.B., 2016. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1525–1533.

Ren, Z., Sudderth, E.B., 2018. 3d object detection with latent support surfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 937–946.

Roberts, L.G., 1963. Machine perception of three-dimensional solids. Ph.D. thesis. Massachusetts Institute of Technology.

Rukhovich, D., Vorontsova, A., Konushin, A., 2021. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. arXiv preprint arXiv:2106.01178.

Song, S., Lichtenberg, S.P., Xiao, J., 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 567–576. doi:10.1109/CVPR.2015.7298655.

Song, S., Xiao, J., 2016. Deep sliding shapes for amodal 3d object detection in rgb-d images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 808–816.

Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W. T., 2018. Pix3d: Dataset and methods for single-image 3d shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A., 2021. Inloc: Indoor visual localization with dense matching and view synthesis. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1293–1307. https://doi.org/10.1109/TPAMI.2019.2952114.

Tan, X., Chen, X., Zhang, G., Ding, J., Lan, X., 2021. Mbdf-net: Multi-branch deep fusion network for 3d object detection. arXiv preprint arXiv:2108.12863.

Tulsiani, S., Gupta, S., Fouhey, D.F., Efros, A.A., Malik, J., 2018. 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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Informat. Process. Syst. 5998–6008.

Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M., 2019. Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7658–7667.

Wang, C., Dai, Y., Elsheimy, N., Wen, C., Retscher, G., Kang, Z., Lingua, A., 2020a. Isprs benchmark on multisensory indoor mapping and positioning. ISPRS Ann. Photogramm. Remote Sens. Spatial Informat. Sci. 5.

Wang, C., Hou, S., Wen, C., Gong, Z., Li, Q., Sun, X., Li, J., 2018. Semantic line framework-based indoor building modeling using backpacked laser scanning point cloud. ISPRS J. Photogramm. Remote Sens. 143, 150–166.

Wang, S., Cai, G., Cheng, M., Junior, J.M., Huang, S., Wang, Z., Su, S., Li, J., 2020b. Robust 3d reconstruction of building surfaces from point clouds based on structural and closed constraints. ISPRS J. Photogramm. Remote Sens. 170, 29–44.

Wang, X., Yeshwanth, C., Nießner, M., 2020c. Sceneformer: Indoor scene generation with transformers. arXiv preprint arXiv:2012.09793.

Xu, D., Anguelov, D., Jain, A., 2018. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 244–253.

Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U., 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Yang, J., Kang, Z., Zeng, L., Akwensi, P.H., Sester, M., 2021. Semantics-guided reconstruction of indoor navigation elements from 3d colorized points. ISPRS J. Photogramm. Remote Sens. 173, 238–261.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Informat. Process. Syst. 32.

Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S., 2021. 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.

Zhang, R., Li, G., Li, M., Wang, L., 2018. Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning. ISPRS J. Photogramm. Remote Sens. 143, 85–96. ISPRS Journal of Photogrammetry and Remote Sensing Theme Issue "Point Cloud Processing".

Zhang, Z., Sun, B., Yang, H., Huang, Q., 2020a. H3dnet: 3d object detection using hybrid geometric primitives. In: European Conference on Computer Vision. Springer.

Zhang, Z., Yang, Z., Ma, C., Luo, L., Huth, A., Vouga, E., Huang, Q., 2020b. Deep generative modeling for scene synthesis via hybrid representations. ACM Trans. Graphics (TOG) 39, 1–21.

Zhou, Y., Zheng, X., Chen, R., Xiong, H., Guo, S., 2018. Image-based localization aided indoor pedestrian trajectory estimation using smartphones. Sensors 18, 258.