Neural Computing and Applications (2023) 35:24407–24433
https://doi.org/10.1007/s00521-023-09025-7

ORIGINAL ARTICLE

Geometric-driven structure recovery from a single omnidirectional image based on planar depth map learning

Ming Meng[1] · Likai Xiao[2] · Zhong Zhou[2,3]

Received: 8 April 2023 / Accepted: 5 September 2023 / Published online: 3 October 2023

Abstract

Scene structure recovery is a crucial process for assisting scene reconstruction and understanding by extracting vital scene structure information and has been widely used in smart city, VR/AR and intelligent robot navigation. Omnidirectional image with a 180° or 360° field of view (FoV) provides greater visual information, making them a significant research topic in computer vision and computational photography. However, indoor omnidirectional scene structure recovery faces challenges like severe occlusion of critical local regions caused by cluttered objects and large nonlinear distortion. To address these limitations, we propose a geometric-driven indoor structure recovery method based on planar depth map learning, aiming to mitigate the interference caused by occlusions in critical local regions. Our approach involves designing an OmniPDMNet, a planar depth map learning network for omnidirectional image, which uses upsampling and a feature-based objective loss function to accurately estimate high-precision planar depth map. Furthermore, we leverage prior knowledge from the omnidirectional depth map and introduce it into the structure recovery network (OmniSRNet) to extract global structural features and enhance the overall quality of structure recovery. We also introduce a distortion-aware module for feature extraction from omnidirectional image, allowing adaptability to omnidirectional geometric distortion and enhancing the performance of both OmniPDMNet and OmniSRNet. Finally, we conduct extensive experiments on omnidirectional dataset focusing on planar depth and structure recovery demonstrate that our proposed method achieves state-of-the-art performance.

Keywords Structure recovery · Omnidirectional image · Planar depth map learning · Distortion-aware learning

1 Introduction

Recovering scene structure from a single image is a fundamental research area in computer vision. It involves inferring the geometry of wall–wall, wall–floor and wall–ceiling boundaries, which serve as essential geometric priors for various applications, including indoor navigation [1, 2], VR/AR/MR [3, 4] and design [5, 6]. Moreover, structure recovery plays a crucial role in scene understanding, aiding object detection and layout recovery [7, 8]. Traditional perspective images captured using the standard pinhole projection model [9–11] have made some progress in scene structure recovery. However, their limited FoV makes it challenging to capture the overall scene structure and context information effectively. In contrast, omnidirectional images with ultra-wide FoV offer richer global contextual information for scene structure recovery.

In this work, our focus lies in indoor structure recovery from omnidirectional image, and existing progress can be classified into three categories. One is the geometry-based method [12, 13], which leverages geometric features to generate structural hypotheses, sort and optimize them, extracting the most reasonable structural recovery. The

✉ Zhong Zhou
zz@buaa.edu.cn

Ming Meng
mengming@cuc.edu.cn

Likai Xiao
xiaolikai@buaa.edu.cn

[1] School of Data Science and Media Intelligence, Communication University of China, Beijing 100024, China

[2] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[3] Zhongguancun Laboratory, Beijing 102629, China

second category is hybrid-driven approaches [14–16], combining geometric features with semantic cues, which have shown improved performance given the remarkable progress of deep neural networks in image semantics. The third category is end-to-end omnidirectional image structure recovery learning frameworks [17, 18], which enhance structure recovery by designing different networks. In the feature extraction module, standard convolution kernels are employed for omnidirectional image feature learning. The trained network predicts relevant elements(structural boundaries or corners) in key indoor structural regions. However, the fixed sampling strategy of standard convolution limits the network's ability to model geometric transformations, making it challenging to address severe geometric distortions in omnidirectional images. Recent studies [18, 19] have explored the distortion of omnidirectional image and proposed introducing deformable convolutions to enhance the learning and modeling capabilities of geometric transformations. Nevertheless, different projection models exhibit distinct positional and distortion characteristics. To tackle this issue, we adopt a distortion-aware module based on the projection model, enabling effective feature extraction from the omnidirectional image.

Moreover, the key geometric structure regions of complex scenes are often partially or completely occluded, posing challenges in achieving high-precision scene structure recovery. Structural features are closely related to depth information. To address this issue, we introduce plane depth into the omnidirectional image structure recovery network, providing strong geometric clues to effectively reduce the interference of occlusions and further enhancing the performance of the structure recovery network to obtain high-quality structure recovery. Conventional depth estimation method from omnidirectional image adopts the perspective split method [20], but it does not fully exploit the global context information and is often time-consuming and inefficient. The subsequent methods based on projection fusion [21, 22] can mitigate distortion to some extent by jointly training omnidirectional and stereo projection maps, but they are difficult to train and have a high time cost. Recently, popular methods focus on optimizing depth estimation from the learning level of omnidirectional features [23, 24] and extracting effective features from omnidirectional image using various types of distortion-aware convolutional filters. However, the depth of the network structure directly impacts the learning capability of the deep learning network, and fuzzy depth estimation at the object edges can become an issue.

In this paper, we propose a geometric-driven network based on planar depth map learning for achieving high-quality structure recovery from omnidirectional indoor image, as shown in Fig. 1. To this end, we construct a new network for planar depth map estimation from omnidirectional image (OmniPDMNet). To enhance depth estimation accuracy, we incorporate the upsampling to deepen the network structure, and we introduce a feature-based loss function to address the issue of object edge ambiguity. Additionally, we devise a structure recovery network from omnidirectional image (OmniSRNet), which utilizes the omnidirectional planar depth map as prior knowledge to extract global and precise structural features. This enables the network to effectively handle challenges posed by severe occlusions, ultimately enhancing the quality of structure recovery. Moreover, to tackle the problem of large space-varying distortion, we introduce a distortion-aware module for both OmniPDMNet and OmniSRNet, thereby improving the performance of learning omnidirectional geometric feature. Finally, we extensively evaluate our proposed method on both virtual and real-world omnidirectional datasets. The experimental results demonstrate the superior performance of OmniSRNet compared to state-of-the-art methods. This affirms the effectiveness and potential of our approach for high-quality indoor structure recovery from omnidirectional image.

To summarize, we discover the correlations between structure recovery and planar depth map learning. Our key contributions are:

- We propose a new omnidirectional structure recovery network driven by geometric prior knowledge, known as a planar depth map from an omnidirectional image. It can significantly alleviate the critical local regions' interference from cluttered objects to generate high-quality omnidirectional structure recovery. Moreover, we demonstrate its flexibility through several applications, including MR video surveillance and VR house viewing.
- We devise a new planar depth map learning network with the upsampling and adopt a feature-based loss function to generate an accurate depth map of the omnidirectional image. Furthermore, we introduce a distortion-aware module into omnidirectional feature extraction, significantly enhancing the network's ability to handle large nonlinear distortion and further improving the performance of depth estimation and structure recovery.
- We conduct extensive omnidirectional structure recovery experiments on both virtual and real-world omnidirectional datasets, encompassing panorama and fisheye images. Our proposed method surpasses state-of-the-art methods in terms of quantitative metrics and visual results, affirming its effectiveness and superiority. The datasets and code will be published at https://github.com/mmlph/OmniSRNet/.

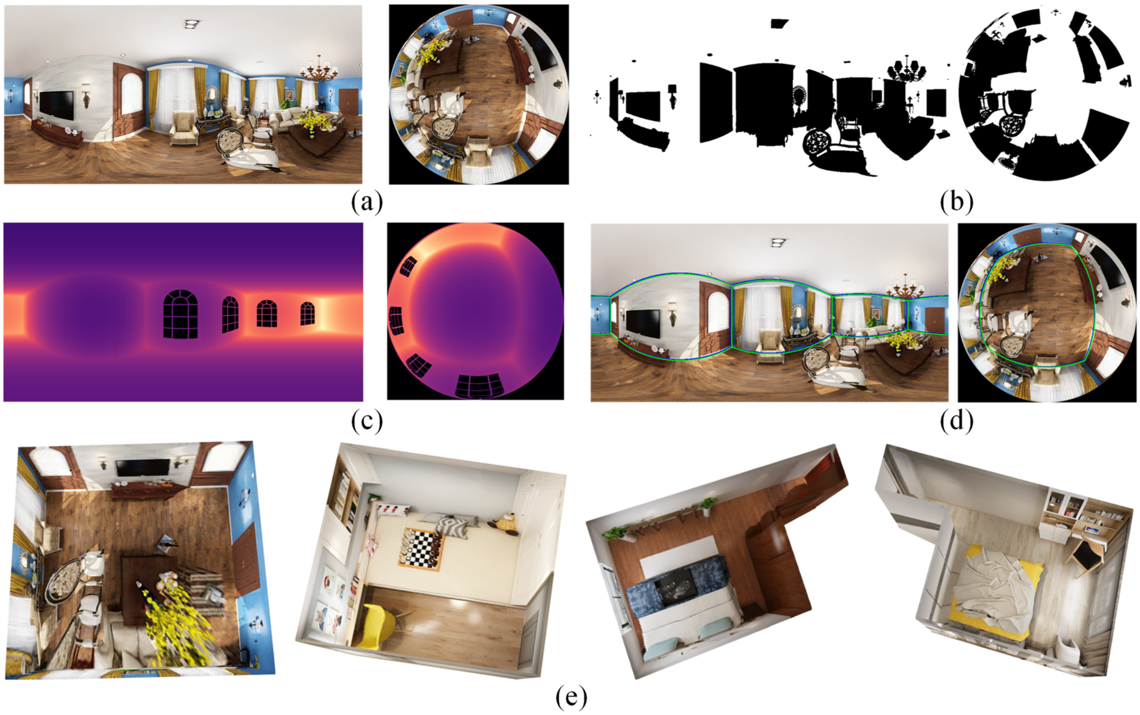

Fig. 1 Given a single omnidirectional image (panorama or fisheye) (**a**) and the corresponding mask image (**a**), our proposed OmniPDM-Net first estimates the planar depth map (**a**). Then our OmniSRNet driven by geometric prior (planar depth map) recovers accurate indoor structure (marked by green), whereas the ground truth is marked by blue (**a**). The corresponding 3D structures by our method, cuboid and non-cuboid, are generated by post-processing (**a**)

2 Related work

We review previous representative approaches on structure recovery and depth estimation from omnidirectional image that are most relevant to our work.

2.1 Structure recovery from the omnidirectional image

The traditional structure recovery methods of perspective images [9–11, 25–28] mainly determine the intersecting boundary between indoor floor and walls under the condition of ‘Manhattan world hypothesis’. Subsequently, they generate three-dimensional structure by these wall boundaries. However, since standard camera can only capture a limited 15% FoV of the human visual system, the corresponding perspective image only records part of the scene structure information, which seriously affects the effect of overall scene structure recovery. While omnidirectional images benefit from their ultra-wide omnidirectional FoV (180° or 360°), which can maximize the coverage of the overall scene information and provide rich global context information for scene structure recovery and understanding. According to different technologies, existing structure recovery approaches are divided into three categories: geometric-based, data-driven and depth-guided methods.

2.1.1 Geometric-based methods

The key idea in *geometric-based* structure recovery is to follow the strategy of ’feature extraction to hypothesis generation to scoring ranking’. The pioneering work of Zhang et al. [12] proposed an indoor scene recovery method from panorama image by making full use of the contextual feature information. However, it needs the process of projection transform splitting, which is time-consuming and inefficient. Yang et al. [13] inferred the panorama depth information by extracting superpixels and line segments and then recovered the indoor structure by combining depth, geometric features and semantic features. Xu et al. [29] combined surface normal vector and object information to recover the scene structure and inferred the position and orientation of objects achieving the overall recovery. Yang et al. [30] designed a structure recovery network for panorama image with a dual encoder–decoder branch (DulaNet) to alleviate occlusion problems. However, the hypothetical optimization strategies of the above methods all bring high computational complexity, and the recovery effect also largely depends on the quality of the extracted semantic information.

2.1.2 Data-driven methods

Inspired by the remarkable performance of Convolutional Neural Networks(CNNs) in feature extraction, *data-driven* structure recovery becomes attractive. One is to alleviate the occlusion problem in the scene by extracting high-quality depth or semantics and using a hybrid geometry data-driven to infer the global optimal spatial structure [14–16]. Subsequently, more and more end-to-end methods have made great strides in omnidirectional structure recovery. The pioneering work of Zou et al. [31, 32] proposed and improved an end-to-end deep convolutional neural network (LayoutNet$_{v1}$ and LayoutNet$_{v2}$) for structure recovery from panorama image. It adopts an encoder–decoder strategy to quickly infer the 3D scene structure. Following this strategy, Sun et al. [33] devised a panorama structure recovery network based on the one-dimensional representation of the structure, which can improve the performance of the network by reducing the model parameters, and can minimize the cost. Jiang et al. [17] proposed a structure recovery network containing Transformer architecture to model geometry relations. Given the serious nonlinear distortion in omnidirectional images, Fernandez et al. [19] proposed a panorama structure recovery network (CFL$_{std}$ and CFL$_{equi}$) based on deformable convolution. And then Rao et al. [18] introduced spherical convolution to replace deformable convolution (OmniLayout).

2.1.3 Depth-guided methods

Depth information is closely related to structure recovery, *depth-guided* scene structure recovery alleviates the object occlusion problem and optimizes the recovery effect by using the depth map as an intermediate representation. Perez-Yus et al. [34] constructed a hybrid camera system combining traditional depth cameras and fisheye cameras. It can combine large viewing angles with depth data and generates corresponding structural hypotheses through the detected structural corners to achieve the overall structure recovery. Zhang et al. [35] inferred the depth maps of the dominant planes in the scene and used the intersection of the depth maps to generate the scene structure from a traditional perspective image. Zeng et al. [36] also adopted the complementary features of geometric structure and depth information and leveraged the depth to reduce the occlusion of the structure by the cluttered objects in the complex scene improving the structure recovery quality (JLDNet). The above methods can alleviate the interference of occlusion on structure recovery to a certain extent, but the representation and quality of depth estimation play a crucial role in the performance of structure recovery.

2.2 Depth estimation from the omnidirectional image

Depth information estimated from a single image is crucial for indoor navigation [37], 3D map reconstruction [38] and 3D scene understanding [39]. Traditional depth estimation methods generate regularized depth maps through non-automated feature selection and probabilistic image models [40, 41], which often suffer from over-constrained scene geometry. Subsequently, researchers focus on multi-scale networks [42–44] and robust loss functions [45] for conditional random fields to continuously improve the depth estimation accuracy from the perspective image.

With the popularity of omnidirectional cameras, depth estimation techniques for omnidirectional images have also been widely used. Researchers mainly adopt omnidirectional feature learning or multiple projection fusion strategies to alleviate the distortion interference of omnidirectional depth estimation. Tateno et al. [46] introduced a panorama image depth estimation method based on a deformable convolution filter correcting the receptive field to achieve more accurate depth estimation. Zioulis et al. [47] proposed a learning-based dense depth estimation network (RectNet) for panorama images. Subsequently, Eder et al. [48] devised a dense depth estimation method for panoramic images based on plane perception. Cheng et al. [23] designed a dilated convolutional depth estimation network. Chen et al. [24] also proposed a distortion-aware dense depth estimation network for panorama image. Wang et al. [21] and Jiang et al. [22] both fully integrated the global field of view of omnidirectional projection with the distortion-free interference feature of stereo projection and designed a dual-branch network with different projection maps as input to further improve the effect of omnidirectional depth estimation. Additionally, Jin et al. [49] proposed a learning-based depth estimation framework, which leverages the geometry structure of the scene as prior knowledge to conduct depth estimation.

Although the above methods have made some progress, the gradient disappearance and overfitting in the depth estimation network model, as well as the blurring of object edges in the depth estimation results, reduce the accuracy of depth estimation of the scene structure and affect the three-dimensional structure of the scene. To provide an effective guarantee for structure recovery, we design a new planar depth map learning network introducing upsampling and adopting a feature-based loss function to generate an accurate depth map of the omnidirectional image.

3 Proposed approach

3.1 System overview

The overall method mainly includes three parts: (1) network architecture for planar depth map learning from omnidirectional image (OmniPDMNet), (2) network architecture with depth-driven for structure recovery from omnidirectional image (OmniSRNet) and (3) post-processing for 3D structure recovery. For the first two parts, an encoding–decoding strategy is used to design the corresponding network architecture, and the planar depth map and the structure corner probability map are estimated, respectively. We obtain the corresponding geometric structure and three-dimensional point cloud through peak processing of a diagonal probability map. The overview of our framework is illustrated in Fig. 2.

The input of our algorithm is an omnidirectional RGB image and the corresponding omnidirectional object mask, and they are jointly input into the planar depth map learning network from an omnidirectional image (OmniPDMNet). The object mask guides the network to predict structural depth maps that only contain depth information on the main planes (ground, wall and ceiling). To improve the accuracy of the estimated depth map, the network uses ResNet50 [50] based on distortion perception in the encoder of feature extraction, which can perform effective omnidirectional feature learning and reduce the interference of omnidirectional distortion on depth estimation.

On this basis, the planar depth map as geometric prior knowledge is introduced into the omnidirectional image structure recovery network (OmniSRNet). This introduction of depth information enables the network to predict a more accurate structural corner probability map, which describes the location of key corner regions where the main planes intersect. The two networks adopt a phased training strategy, in which the planar depth estimation network uses a gradient-based loss function for network convergence to solve the problem of object edge ambiguity, while the structure recovery network uses the binary cross-entropy loss function to converge the network and improve the accuracy of the predicted corner probability map. Finally, we recover the 2D structure of the ground, wall and ceiling in the omnidirectional image, and 3D structure through an omnidirectional image 3D point cloud recovery method.

3.2 OmniPDMNet: network architecture for planar depth map learning from the omnidirectional image

For an empty scene without occlusion, the depth of field of the scene has a strong correlation with the geometric structure. The corner points of the scene structure are located at the local maximum depth of field, and the depth of field distribution in the same straight line or the same plane presents a regular pattern. Therefore, we adopt an encoder–decoder strategy and devise a planar depth estimation network for omnidirectional image to predict the planar depth map removed movable objects

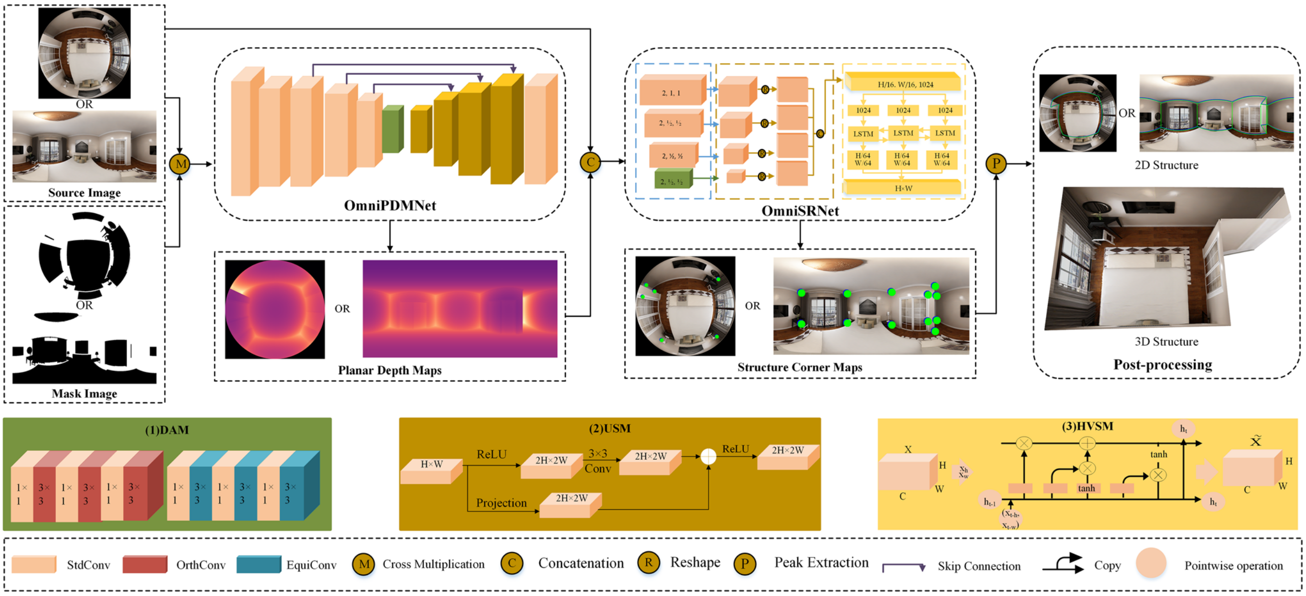

Fig. 2 Overview of our approach for room structure recovery from a single omnidirectional image. Given an input omnidirectional image (fisheye or panorama) and the corresponding mask image, our approach outputs the bounding cuboid (convex polyhedra) or general non-cuboid representation of the indoor scene. It consists of (1) the planar depth map learning network for omnidirectional image (OmniPDMNet) and (2) the structure recovery network driven by planar depth map for omnidirectional image (OmniSRNet)

Table 1 Ablation study for depth estimation on omnidirectional datasets (panorama and fisheye) with variations in Conv

Dataset	Model		Error↓			Accuracy↑		
	Aspect	Description	ABS_REL	SQ_REL	RMSE	δ_1	δ_2	δ_3
Panorama	Conv.Type	OmniPDMNet w/ StdConv[33]	0.478	1.779	0.49	0.928	0.969	0.977
		OmniPDMNet w/ DCNv1[55]	0.453	1.559	0.483	0.929	0.97	0.979
		OmniPDMNet w/ DCNv2[56]	0.448	1.467	0.471	0.931	0.975	0.981
		OmniPDMNet w/ EquiConv	**0.414**	**1.223**	**0.421**	**0.932**	**0.978**	**0.981**
	Loss.Type	OmniPDMNet w/ EquiConv+Huber[57]	0.411	1.206	0.419	0.938	0.981	0.984
		OmniPDMNet w/ EquiConv+Feature	0.312	1.099	0.314	0.949	0.984	0.990
	Mask.Strategy	OmniPDMNet w/ EquiConv+Feature+Encoder	**0.298**	**0.976**	**0.213**	**0.952**	**0.987**	**0.995**
		OmniPDMNet w/ EquiConv+Feature+Decoder	0.309	0.998	0.307	0.951	0.986	0.992
Fisheye	Conv.Type	OmniPDMNet w/ StdConv[52]	0.567	2.046	0.763	0.916	0.956	0.961
		OmniPDMNet w/ DCNv1[55]	0.561	1.983	0.718	0.916	0.958	0.961
		OmniPDMNet w/ DCNv2[56]	0.487	1.659	0.663	0.919	0.962	0.971
		OmniPDMNet w/ OrthConv	**0.430**	**1.353**	**0.581**	**0.925**	**0.964**	**0.972**
	Loss.Type	OmniPDMNet w/ OrthConv+Huber[57]	0.422	1.328	0.565	0.928	0.971	0.980
		OmniPDMNet w/ OrthConv+Feature	0.364	1.238	0.383	0.936	0.979	0.988
	Mask.Strategy	OmniPDMNet w/ OrthConv+Feature+Encoder	**0.306**	**1.013**	**0.335**	**0.948**	**0.981**	**0.990**
		OmniPDMNet w/ OrthConv+Feature+Decoder	0.347	1.202	0.375	0.944	0.980	0.988

Type (different convolutions), Loss.Type (Huber and Feature) and Mask.Strategy (Encoder and Decoder). The error and accuracy are shown in %, and the results are highlighted with bold numbers in blue, yellow and green, indicating relatively superior performances for each ablation study. The color green corresponds to the best overall performance achieved by OmniPDMNet. Evaluation metrics with (↓) signify smaller values being better, while metrics with (↑) indicate larger values being preferable

(OmniPDMNet). Taking it as geometric prior knowledge to guide the structure recovery of complex scenes, which can minimize the impact of clutter on key structural areas, and obtain high-quality structure recovery. To improve the accuracy of the structure depth, we introduce the distortion-aware module (DAM) into the encoder to extract accurate omnidirectional features. And the upward mapping layer module based on the residual idea is added to the decoder improving the network learning ability. What's more, we design a feature-based loss function to optimize the depth estimation at the object edge. The overview of the OmniPDMNet is illustrated in Fig. 2, where we introduce several important novelties in detail as follows.

3.2.1 Network input

The input of OmniPDMNet consists of two parts, the omnidirectional RGB image and the corresponding mask image. The omnidirectional image includes panorama image with the standard equirectangular projection (ERP) and fisheye image with general orthographic projection (Orth). The resolution of the input omnidirectional image, 3 × H × W (for channel, height and width), is a hyper-parameter and can be adjusted according to the experiment condition. The mask image is a bitmap, and all the pixels corresponding to the movable object are set to 0 and presented in black. The pixel value of other structural areas is set to 255 and presented in white.

3.2.2 Encoder with DAM

In our approach, we leverage ResNet50 as the backbone network to extract relevant low/mid/high-level semantic features from the input omnidirectional image. However, the omnidirectional image does not conform to the pinhole camera model [9, 11, 51], and the distortion cannot be described by the conventional perspective relationship. The standard convolution with fixed sampling strategy, commonly employed in convolutional neural networks, restricts the receptive field of feature expression and limits the network's ability to model geometric transformations, making it difficult to perform effective omnidirectional feature learning. Following [52], we introduce the distortion-aware module (DAM) into the last convolutional layer of ResNet50 to improve the modeling ability of geometric transformation in structural depth estimation. Specially, we replace the standard convolution (of 3 × 3 filters) with EquiConv [19] or OrthConv [52] in the last block for panorama image and fisheye image, respectively.

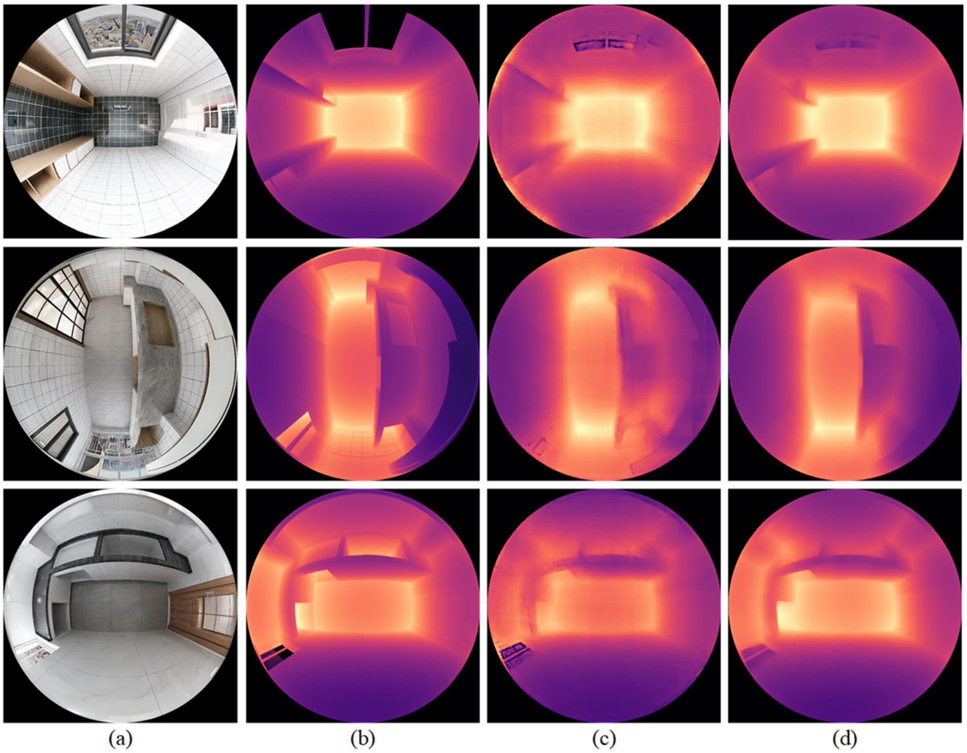

Fig. 3 Qualitative comparison results of depth estimation between different loss functions on fisheye dataset. Left to Right: For each fisheye image (**a**), we show its depth estimating result by OmniPDMNet using Huber-based (**c**) and our Feature-based (**d**) loss function, whereas the ground truth is (**b**)

Table 2 Quantitative comparison for depth estimation on omnidirectional datasets (panorama and fisheye) using different networks (FCRN [58] and JLDNet [36])

Dataset	Method	Error↓			Accuracy↑		
		ABS_REL	*SQ_REL*	*RMSE*	δ_1	δ_2	δ_3
Panorama	FCRN [58]	0.641	1.205	0.469	0.824	0.954	0.988
	JLDNet [36]	0.524	1.119	0.388	0.889	0.974	0.989
	OmniPDMNet(Ours)	**0.298**	**0.976**	**0.213**	**0.952**	**0.987**	**0.995**
Fisheye	T-FCRN	0.421	1.416	0.834	0.736	0.941	0.984
	T-JLDNet	0.401	1.302	0.815	0.744	0.962	0.986
	OmniPDMNet(Ours)	**0.306**	**1.013**	**0.335**	**0.948**	**0.981**	**0.990**

The error and accuracy are shown in % and bold numbers indicate the best performance. For evaluation metrics with (↓), smaller is better, while for evaluation metrics with (↑), bigger is better

3.2.3 Decoder with USM

The depth estimation network constructed with conventional decoders will lead to gradient disappearance and overfitting phenomenon due to the less depth of the network layer, which weakens the ability of omnidirectional structure depth learning. To this end, we design a decoder consisting of four upsampling modules (USM) and a 3 × 3 convolutional layer. We use bilinear interpolation for upsampling to increase the resolution of the feature map to be consistent with the original image. The USM based on the residual structure can further increase the depth of the network structure, avoid gradient disappearance and overfitting problems and improve the learning ability of the depth estimation model. Additionally, we fuse the multi-scale features in the encoder and decoder with skip connections. This can fully utilize the omnidirectional semantic information in the feature maps of different scales and further improves the accuracy of the structural estimation depth.

3.2.4 Feature-based loss function

Depth estimation is a typical regression task, and the mean square error loss function (mean squared error loss, MSE) is usually used to optimize the regression problem during the training process of the estimation network model. Under the assumption that the collected samples obey the same Gaussian distribution, MSE uses the residual term as the loss value in a direct way. However, the image captured

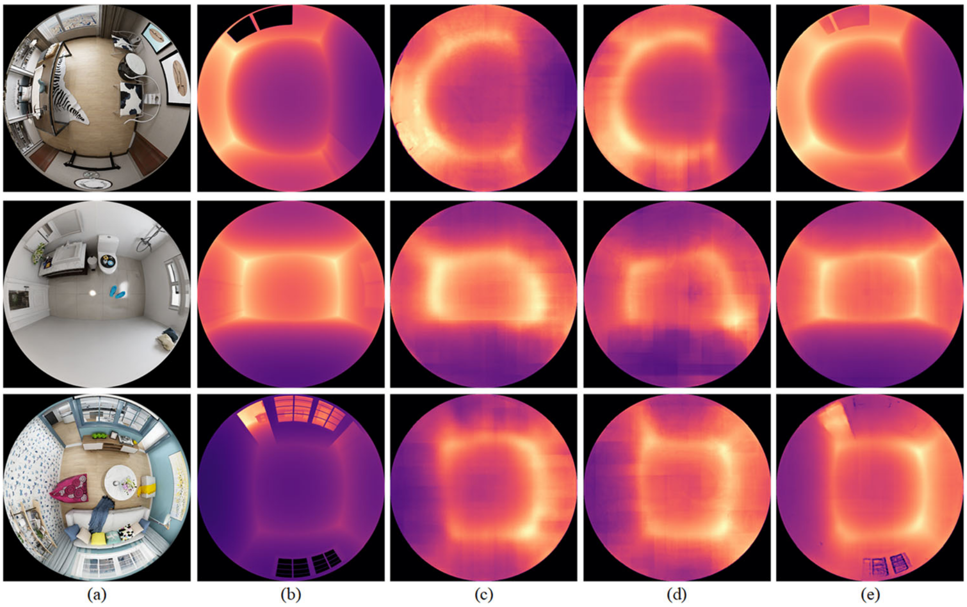

Fig. 4 Qualitative comparison of different depth estimation networks on fisheye dataset. Left to Right: For each fisheye image (**a**), we show its planar depth map estimated by (**c**) T-FCRN, (**d**) T-JLDNet and (**e**) our OmniPDMNet, respectively, whereas the ground truth is (**b**)

Table 3 Quantitative comparison for structure recovery on omnidirectional datasets (Synthetic and Real-world) without or with planar depth map

Dataset	Depth	Train→ Finetune	Panorama			Fisheye		
			CE (%)↓	PE (%)↓	3DIoU (%)↑	CE (%)↓	PE (%)↓	2DIoU (%)↑
Synthetic	w/o	w/o	0.25	0.69	96.31	0.67	0.55	95.88
	w/	w/o	**0.22**	**0.68**	**98.44**	**0.39**	**0.55**	**97.94**
Real-world	w/o	w/o	1.86	2.65	80.62	4.54	2.04	76.65
	w/	w/o	0.53	1.50	87.81	2.54	1.06	87.16
	w/	w/	**0.50**	**1.00**	**88.81**	**2.02**	**0.92**	**88.2**

The accuracy is shown in %, and bold numbers indicate the best performance. w/o and w/ finetune indicate whether to use synthetic dataset for pre-training. For evaluation metrics with (↓), smaller is better, while for evaluation metrics with (↑), bigger is better

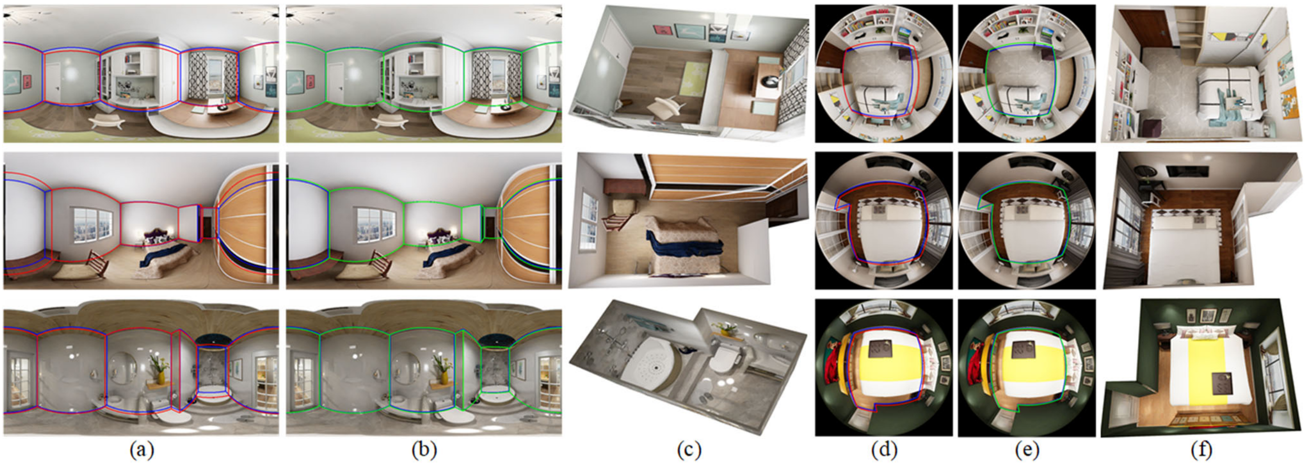

Fig. 5 Qualitative comparison of the effect of planar depth map for structure recovery on synthetic datasets (**a** ∼(**c**) for panorama dataset, **d** ∼(**f**) for fisheye dataset). Left to Right: For each omnidirectional image, we show its structure recovered by (**a**, **d**) w/o depth-driven, (**b**,**e**) w/ depth-driven, (**c**,**f**) point cloud processing. The predictions of our method with and without depth-driven are highlighted in green and red, respectively, whereas the ground truth is in blue

from real-world scenes contains many features that make it difficult to obey the uni-modal Gaussian distribution. To this end, we design the target loss function according to the loss terms corresponding to different features in the image, which mainly include depth estimation loss term, gradient prediction loss term and surface normal loss term.

3.2.5 Depth estimation loss term

The most common feature in actual scene image is that the objects are different in distance. If this feature is ignored and using the depth error value of the same weight to directly calculate the depth loss term, it will cause blurring in the depth estimation result. Therefore, we design the depth error with different weight proportions according to the distance between the object and the camera, specifically as follows:

$$L_{depth} = \frac{1}{N}\sum_{i=1}^{N} \ln(\|d_i - g_i\| + \alpha) \tag{1}$$

where α is an adjustable parameter, and the empirical value is set to 0.5. This term uses the logarithmic form of the depth error instead of the commonly residual term, making the distance between the object and the camera proportional to the weight. The closer the object is to the camera, the greater the error proportion will be. On the contrary, the farther the object is, the smaller the weight will be. The trained network model is more stable and beneficial for structural depth estimation.

Gradient prediction loss term. Another feature that appears most in the image is the multistep structure of edges. The depth estimation loss term only balances the depth transformation direction, and it is difficult to deal with different offsets in the depth direction. To solve the problem of the blurred phenomenon at the edge of the object, we first adopt the Sobel gradient operator to extract the edge of the feature map. After that, we introduce the gradient prediction loss term in the process of neural network back-propagation and the expression as follows:

$$L_{grad} = \frac{\sum_{i=1}^{N}\left(\ln(\nabla_x(\|d_i - g_i\|) + \alpha) + \ln\left(\nabla_y(\|d_i - g_i\|) + \alpha\right)\right)}{N} \tag{2}$$

where ∇_x and ∇_y are the gradient magnitudes represented by the vector, which represent the partial derivatives of the depth error in the x and y directions, respectively.

3.2.6 Surface normal loss term

Although the gradient loss mentioned above can optimize the different depths of the edge, it is difficult to effectively deal with the shape features in the scene, such as the common main shape features (steep edges, corners and plane structures). The normal vector can encode the angle information of the surface of the object, and the plane feature can be globally constrained by a unified normal vector. At the same time, the angle information can also be used to effectively constrain the local structural features. Therefore, we introduce the surface normal loss term in the loss function and use the constraint of surface normal to improve the estimation accuracy of the global and local details of the structural depth map. It is specifically expressed as:

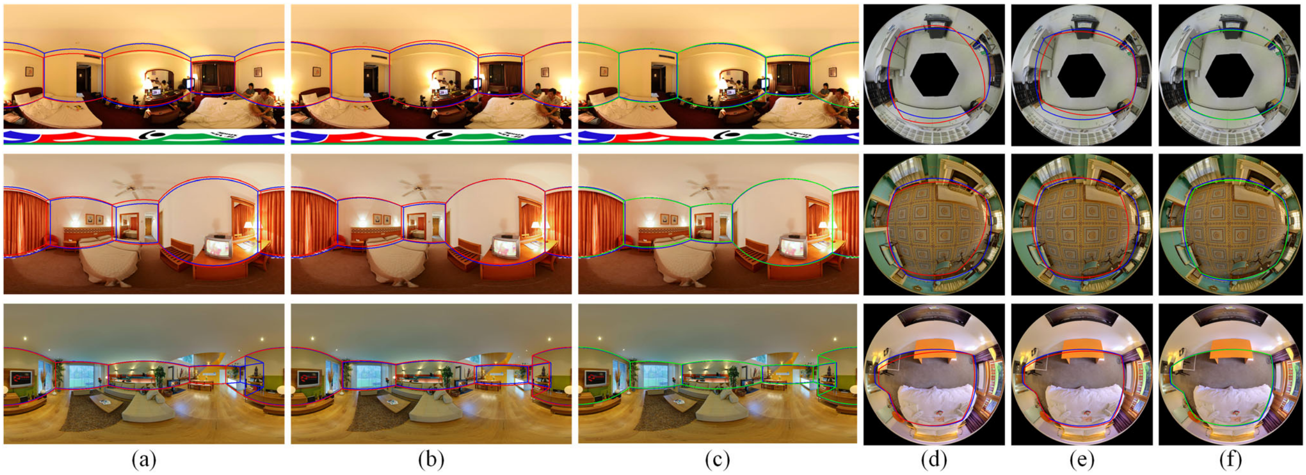

Fig. 6 Qualitative comparison of the effect of planar depth map for structure recovery on real-world datasets (**a** –**c**) for panorama dataset, **d**–**f** for fisheye dataset). Left to Right: For each omnidirectional image, we show its structure recovered by OmniSRNet without fine-tuning and without depth-driven (**a**, **d**), with fine-tuning but without depth-driven (**b**, **e**), with fine-tuning and with depth-driven (**c**, **f**). The predictions of fine-tuning strategies with depth and others are highlighted in green and red, respectively, whereas the ground truth is in blue

Table 4 Quantitative comparison of [17–19, 31–33, 36] and our approach on our refined panorama datasets

Methods	PanoContext			Stanford2D3D			Structured3D		
	CE (%)↓	PE (%)↓	3DIoU (%)↑	CE (%)↓	PE (%)↓	3DIoU (%)↑	CE (%)↓	PE (%)↓	3DIoU (%) ↑
LayoutNet$_{v1}$ [31]	1.02	3.81	71.42	0.92	2.42	77.51	1.44	2.98	82.86
LayoutNet$_{v2}$ [32]	0.93	2.81	76.90	0.88	2.78	78.90	1.35	2.87	83.24
CFL$_{std}$ [19]	0.79	2.49	78.79	1.44	4.75	65.13	1.87	3.97	78.91
CFL$_{equi}$ [19]	0.78	2.64	77.63	1.64	5.52	65.23	2.04	4.39	78.19
HorizonNet [33]	0.76	2.13	83.39	0.63	2.06	84.09	0.31	1.01	93.74
OmniLayout [18]	0.69	2.10	84.50	0.68	2.14	83.40	0.31	2.73	93.88
JLDNet [36]	0.71	2.08	86.21	0.61	1.74	84.44	0.29	0.98	95.42
LGTNet [17]	0.69	2.07	85.16	0.63	2.11	85.76	0.29	0.84	96.21
OmniSRNet(Ours)	**0.48**	**0.93**	**88.64**	**0.52**	**1.06**	**88.97**	**0.22**	**0.68**	**98.44**

The accuracy is shown in %, and bold numbers indicate the best performance. For evaluation metrics with ↓, smaller is better, while for evaluation metrics with ↑, bigger is better

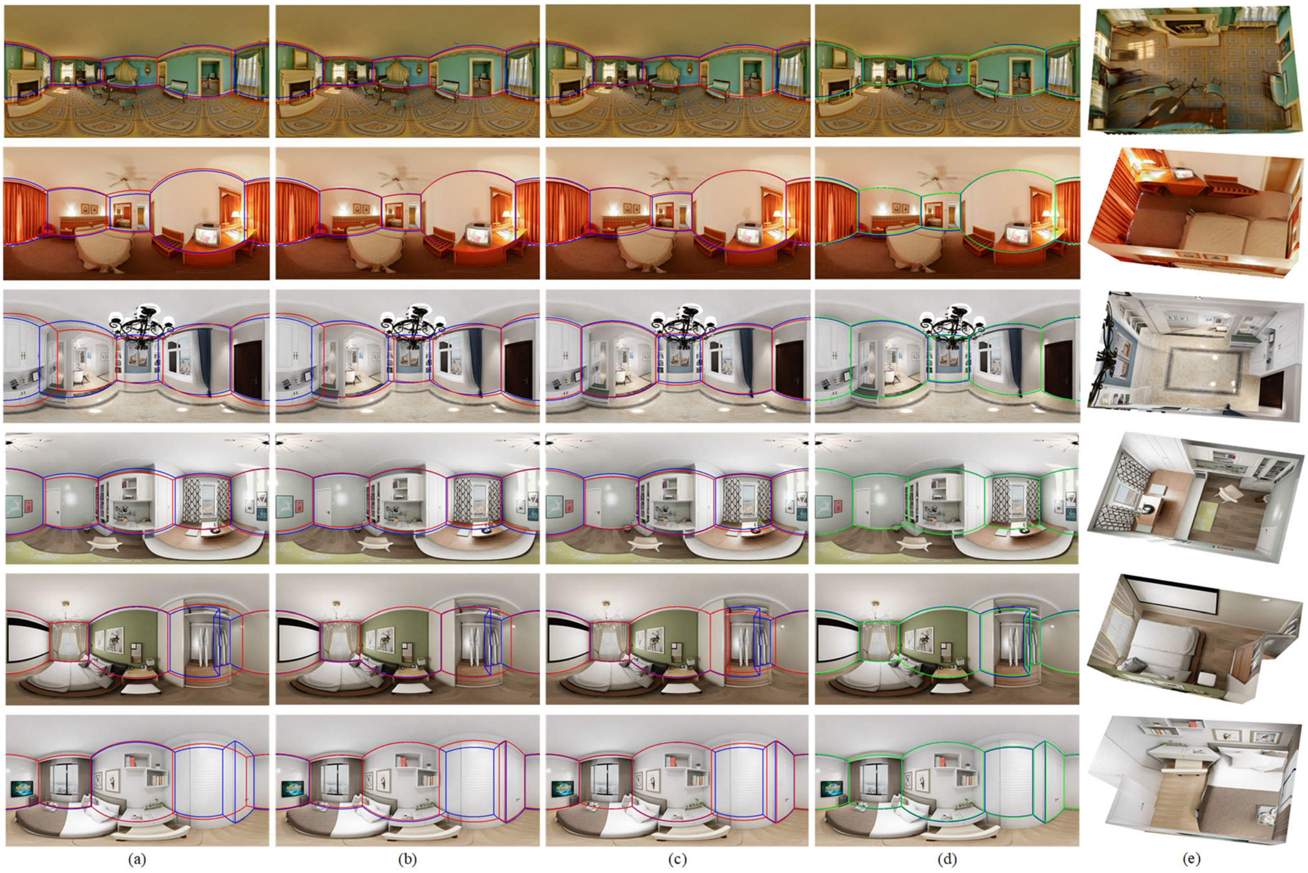

Fig. 7 Qualitative comparison of different methods on panorama datasets. Left to Right: For each panorama image, we show its structure recovered by (**a**) HorizonNet [33], **b** JLDNet [36], **c** LGTNet [17], **d** our method with planar depth-driven and **e** 3D point cloud. The predictions of our method and others are highlighted in green and red, respectively, whereas the ground truth is in blue

Table 5 Quantitative comparison of the modified version of [17–19, 31–33, 36] and our approach on our fisheye datasets

Methods	PanoContext-F			Stanford2D3D-F			Structured3D-F		
	CE (%)↓	PE (%)↓	2DIoU (%)↑	CE (%)↓	PE (%)↓	2DIoU (%)↑	CE (%)↓	PE (%)↓	2DIoU (%)↑
T-LayoutNet$_{v1}$	7.62	7.20	60.75	8.65	6.71	56.72	5.9	6.23	58.21
T-LayoutNet$_{v2}$	5.73	5.96	75.49	6.64	5.91	70.96	1.33	2.00	90.67
T-CFL$_{std}$	8.53	2.19	62.71	9.89	2.78	57.78	5.96	1.34	68.95
T-CFL$_{equi}$	13.85	3.29	46.21	14.18	3.43	44.82	6.11	1.46	66.79
T-HorizonNet	5.04	2.78	76.32	6.48	3.26	71.04	1.29	1.36	90.67
T-OmniLayout	4.96	2.41	77.64	5.88	3.09	72.49	1.07	1.14	91.92
T-JLDNet	4.36	2.64	82.34	4.18	2.07	80.41	0.88	0.97	94.31
T-LGTNet	4.22	1.97	80.67	4.31	2.29	76.48	0.93	1.03	92.47
OmniSRNet(Ours)	**1.96**	**0.98**	**89.42**	**2.07**	**0.86**	**86.98**	**0.39**	**0.55**	**97.94**

The accuracy is shown in %, and bold numbers indicate the best performance. For evaluation metrics with ↓, smaller is better, while for evaluation metrics with ↑, bigger is better

$$L_{normal} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\langle n_i^d, n_i^g \rangle}{\sqrt{n_i^d, n_i^d}\sqrt{n_i^d, n_i^d}}\right) \quad (3)$$

where $n_i^d = \left[\nabla_x(d_i), \nabla_y(d_i), 1\right]^T$ represents the normal vectors calculated in the predicted depth map. $n_i^g = \left[-\nabla_x(g_i), -\nabla_y(g_i), 1\right]^T$ denotes the ground truth. $\langle n_i^d, n_i^g \rangle$ represents the inner product operation of the predicted normal vector and the ground truth normal vector.

The total feature-based loss function is defined as:

$$L = \omega_1 L_{depth} + \omega_2 L_{grad} + \omega_3 L_{normal} \quad (4)$$

where L_{depth}, L_{grad} and L_{normal} represent the depth estimation loss term, gradient prediction loss term and surface normal loss term, respectively. The overall loss is the sum of the weights of the three losses, and ω_1, ω_2 and ω_3 weight corresponding terms, controlling their importance. In our experiments, we empirically set them to 1.

3.3 OmniSRNet: network architecture with depth-driven for structure recovery from the omnidirectional image

The cluttered arrangement of objects will partially or completely occlude key areas (edges and corners) in the scene structure, making it difficult to extract global structural information, especially in complex Manhattan-type scenes. How to effectively deal with the occlusion phenomenon is the key to high-quality structure recovery. Depth of field has a strong correlation with geometric structure. Based on the study of planar depth, we construct a network that uses depth as a geometric prior to drive structure recovery for omnidirectional image (OmniSRNet) to achieve high-precision results. The overview of the OmniSRNet is illustrated in Fig. 2, where we introduce several important novelties in detail as follows.

3.3.1 Encoder-to-decoder

The raw omnidirectional image and the corresponding planar depth map are fed into a ResNet-based encoder to extract the effective structure features. Similar to [52], we set the corresponding distortion perception modules according to different projection models in the last block of ResNet50 to improve the accuracy of omnidirectional image feature extraction. Furthermore, to capture both low-level and high-level features, the last four feature maps of the encoder are preserved through a series of convolutional layers, and the feature maps are reshaped to the same size, concatenated as a single sequential feature map for Bi-LSTM input. Bi-LSTM is adopted in the decoder to capture long-range geometric patterns of objects for global coherent prediction. A parallel horizontal–vertical stepping module (HVSM) is designed to fully utilize contextual priors of omnidirectional images to achieve high-quality probability map of the predicted indoor corners.

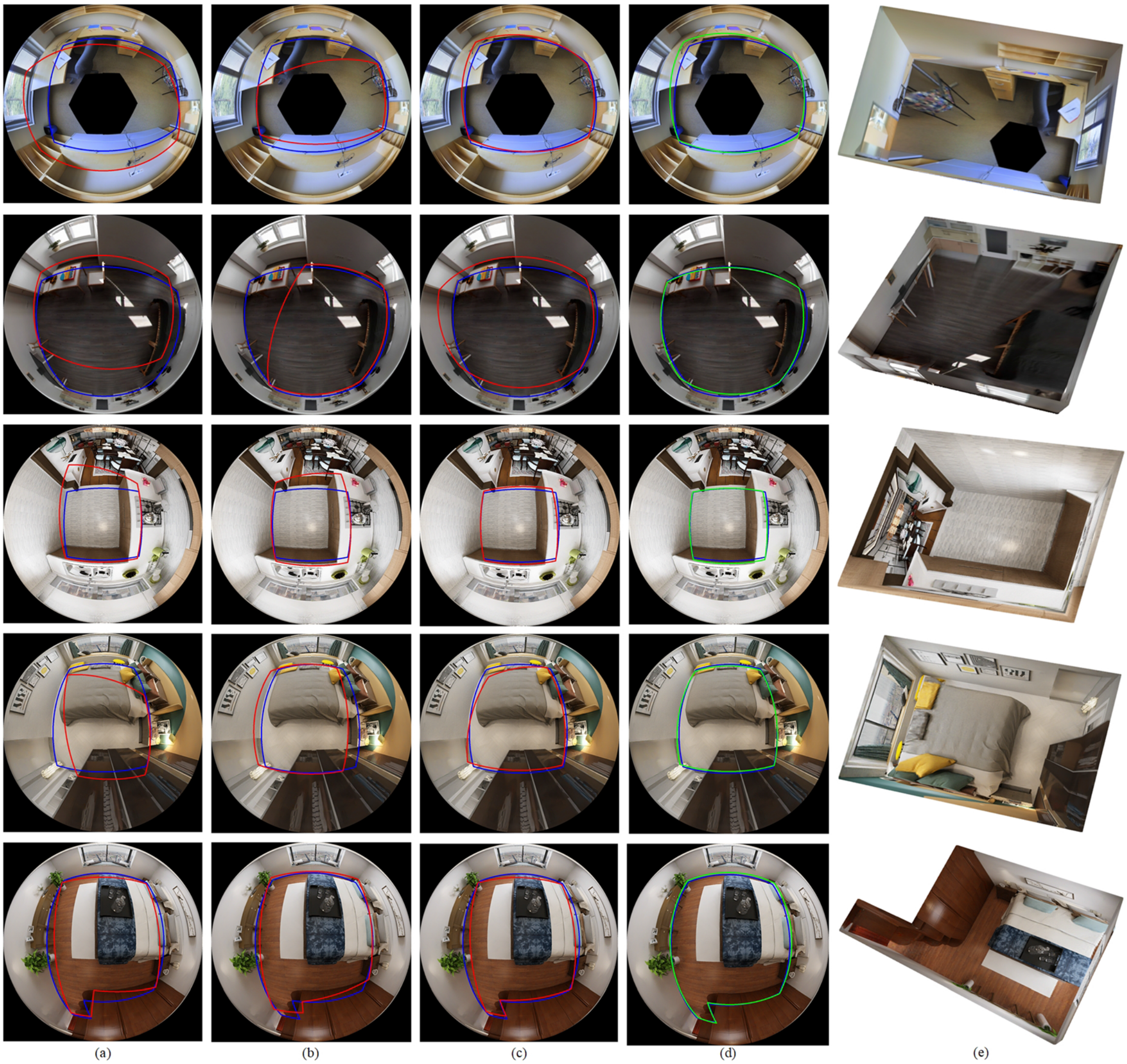

Fig. 8 Qualitative comparison of different methods on fisheye datasets. Left to Right: For each fisheye image, we show its structure recovered by (**a**) T-HorizonNet, **b** T-JLDNet, **c** T-LGTNet, **d** our method with planar depth-driven and **e** 3D point cloud. The predictions of our method and others are highlighted in green and red, respectively, whereas the ground truth is in blue

3.3.2 3D structure recovery

Taking the omnidirectional image and the corresponding mask image as input, the above-mentioned planar depth-guided structure recovery network can predict the probability map of indoor structure corners. Then we post-process the predicted corner probability maps by peak extraction. Under the Manhattan world assumption, the 3D point cloud of the indoor structure is recovered according to the geometric constraints of different projection models. The overall procedure of our method is presented in Algorithm 1.

Algorithm 1 Process of the structure recovery network with depth driven for omnidirectional image.

```
Input: $\mathbf{I_o}$ (an input omnidirectional image, e.g. panorama image or fisheye image);
$\mathbf{I_m}$ (an input mask image, e.g. panorama mask image or fisheye mask image).
Output: 3D point cloud with texture, PCL.
Step 1-Planar Depth Map Estimation
    $\mathbf{E_{pdm}} \leftarrow OmniPDMNet(\mathbf{I_o}, \mathbf{I_m})$
Step 2-Corner Probability Map Prediction
    $\mathbf{P_{cpm}} \leftarrow OmniSRNet(\mathbf{I_o}, \mathbf{E_{pdm}})$
Step 3-Point Structure Recovery
  (a)-Extraction($\mathbf{E_{cpm}}$)
      $\mathbf{E_{cpm}} \leftarrow extend(\mathbf{P_{cpm}})$
      $\mathbf{B_{cpm}} \leftarrow binarize(\mathbf{E_{cpm}})$
      $\mathbf{R_{cpm}} \leftarrow regions(\mathbf{B_{cpm}})$
      $\mathbf{p_c^i} \leftarrow gravity(\mathbf{R^i_{m_c}})$ for $i = 1, ..., N$
  (b)-Recovery($\mathbf{I_o}, \mathbf{p_c}$)
      $\mathbf{T} \leftarrow normalize(\mathbf{I_o})$
      $\mathbf{P_c} \leftarrow coor2xyz(\mathbf{p_c})$
      $\mathbf{z_f} \leftarrow sum(\mathbf{z_i})$ $\mathbf{z_i} \in \mathbf{P_c^i}$
      $\mathbf{P_{plane}} \leftarrow calXYZ(\mathbf{P_c^i}, \mathbf{z_f}, \mathbf{p_n}, \mathbf{p_s})$
      $\mathbf{UV_{plane}} \leftarrow mapUV(\mathbf{P_{plane}})$
      $\mathbf{T_{plane}} \leftarrow interT(\mathbf{UV_{plane}}, \mathbf{T})$
    $\mathbf{PCL} \leftarrow genStructure(\mathbf{UV_{plane}}, \mathbf{T_{plane}})$
```

4 Experimental results

We first briefly introduce the implementation details (Sect. 4.1), containing omnidirectional datasets, the common evaluation metrics and training strategy. For performance evaluation, we validate the effectiveness of planar depth estimation and structure recovery from omnidirectional image(Sects. 4.2 and 4.3), respectively.

4.1 Implementation details

4.1.1 Omnidirectional datasets

We carry out experiments on a large-scale indoor omnidirectional RGB dataset, including panorama datasets (PanoContext [12], Stanford2D3D [53] and Structured3D [54]) and fisheye datasets (PanoContext-F, Stanford2D3D-F and Structured3D-F) constructed by Meng et al. [52]. Both PanoContext and Stanford2D3D are captured from real-world scenes, containing 512 and 550 panorama images with corresponding corner annotations, respectively. While Structured3D is rendered with synthetic scenes, it includes 21521 panorama images with corresponding corner and planar depth map annotations. The same is true for the distribution of the fisheye datasets.

4.1.2 Evaluation metrics

To objectively evaluate the performance of planar depth estimation, we use standard evaluation protocols [21–24, 36, 49] with error metrics and accuracy metrics. The error metrics include ABS_REL (absolute relative error), SQ_REL (square relative error) and RMSE (root mean square error). The accuracy metrics describe the percentage of estimated accurate pixels in all pixels and divided into δ_1, δ_2 and δ_3 according to different thresholds. During the evaluation of structure recovery, we adopt four widely used quantitative metrics used in previous works [17–19, 31, 33], including corner error (CE), pixel error (PE), 2D intersection over union (2DIoU) and 3D intersection over union (3DIoU).

4.1.3 Training strategy

We implement our network (OmniPDMNet and OmniSRNet) on PyTorch platform and train the model on a single RTX 3090 GPU with 24GB. The input size of panorama RGB image and the corresponding ground truth are 1024 × 512, and the size of fisheye image is 1024 × 1024. During the training process, the Adam optimizer is used to update the network parameters, and the maximum learning rate is set to 0.0001. To prevent overfitting during training, we use

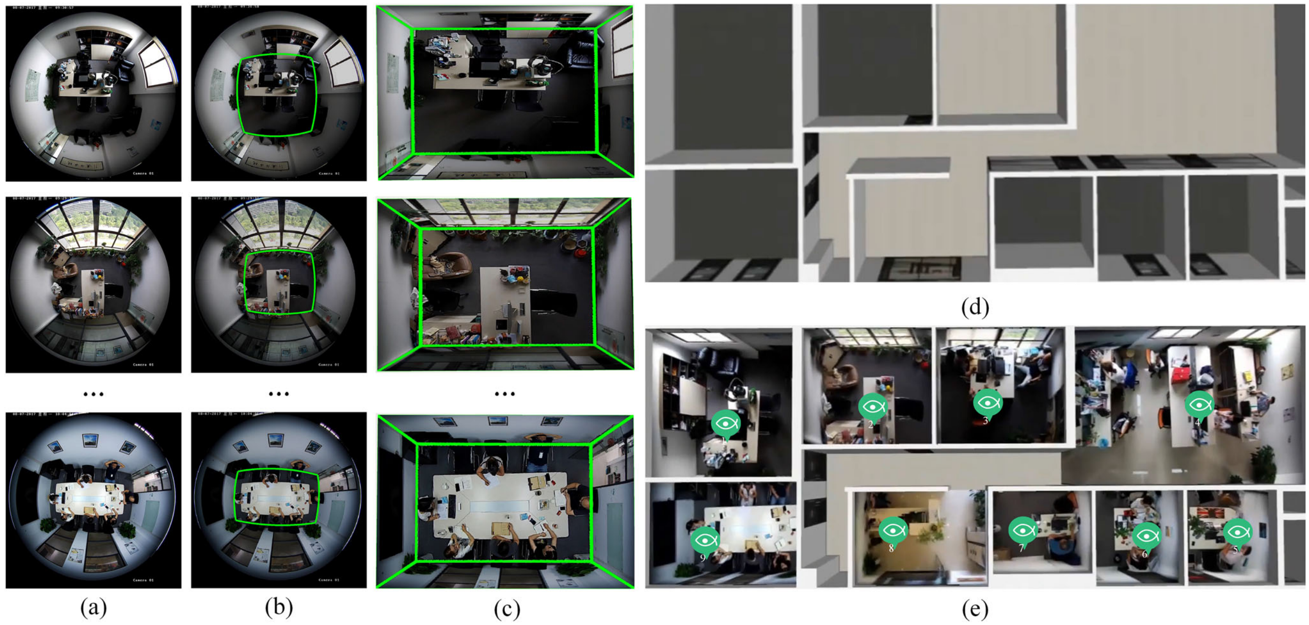

Fig. 9 MR-based video surveillance: a real office environment with nine fisheye images captured by fisheye cameras from a building video surveillance system. **a** The omnidirectional RGB images with 180° FoV. **b** The structure recovery result of our OmniSRNet for each fisheye image (marked by green lines). **c** The 3D structure represented by 3D point cloud. **d** The existing CAD models of the office environment. **e** The texture model recovered from each fisheye image consists of one floor and four walls

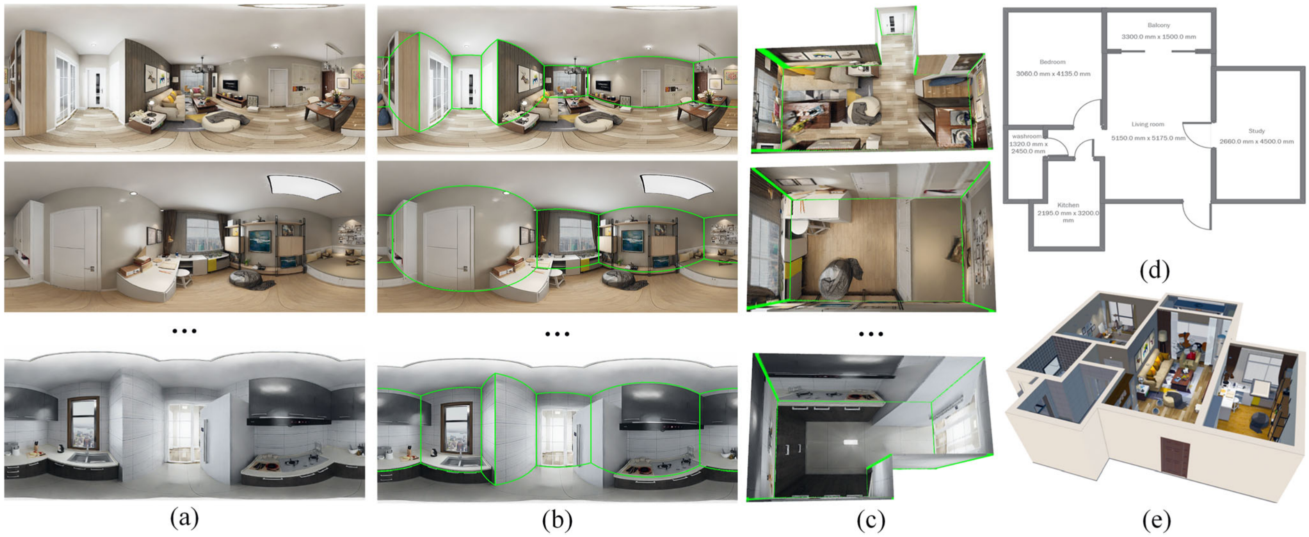

Fig. 10 VR-based house viewing: an interior scene with six panorama images captured by a panorama camera from a real estate. **a** The omnidirectional RGB images with 360° FoV. **b** The structure recovery result of our OmniSRNet for each panorama image (marked by green lines). **c** The 3D structure represented by 3D point cloud. **d** The existing 2D floor plan of a house. **e** The panoramic display from each panorama image consists of floor and walls

L2 regularization to constrain the network parameters. We train OmniPDMNet on virtual omnidirectional datasets with planar depth map annotations. Meanwhile, OmniSRNet is trained on both virtual and real-world omnidirectional datasets with corner annotations.

4.2 Performance evaluation of omnidirectional depth estimation

We perform a series of ablation studies to evaluate the benefit of our introduced modules, including the distortion-aware module, loss function and the mask introduction. And we also validate the effect of our depth estimation network (OmniPDMNet) by comparing with other methods.

4.2.1 Effect of distortion-aware module

To illustrate the effect of distortion-aware module, we compare the omnidirectional depth estimation using different convolutions on both datasets. The quantitative comparison results are shown in Table 1 (Conv.Type). Obviously, the results of panorama depth estimation based on EquiConv are higher than other convolutions (StdConv, DCNv1 and DCNv2) in various quantitative metrics. And compare with StdConv, the SQ_REL error is reduced by nearly 6%, and the accuracy of the network can reach to 98.1%. Also on the fisheye dataset, the OrthConv designed according to the orthogonal projection model can obtain lower depth estimation error and higher depth estimation accuracy. All of these demonstrate that distortion-aware can improve the accuracy of feature extraction and enhance the modeling ability of the depth estimation network for geometric distortion.

4.2.2 Effect of loss function module

To verify the effectiveness of the depth estimation based on our proposed feature loss function, a comparative experiment between different loss functions is carried out on the omnidirectional datasets, and quantitative comparisons are made on error and accuracy indicators, respectively. As illustrated in Table 1 (Loss.Type), our OmniPDMNet with feature-based loss function achieves a certain improvement on panorama and fisheye datasets (0.314 and 0.383 in RMSE, 0.99 and 0.98 in δ_3) compared to Huber loss function.

Figure 3 shows the qualitative comparisons of various loss functions on fisheye dataset. Each column from (a) to (d) presents the original RGB image, ground truth and depth estimation results using Huber-based and Feature-based loss functions, respectively. Obviously, OmniPDMNet with Feature-based loss function is superior to Huber-based, the prediction results at the object edge are more accurate and closer to ground truth, demonstrating that it can better optimize the edges of cluttered objects and obtain high-quality depth estimation results.

4.2.3 Effect of mask introduction module

To alleviate the interference of cluttered occlusions, the structure recovery is driven by the planar depth map as the geometric prior knowledge, while it is generated in the depth estimation network guided by the object mask map. The mask map can be introduced in different strategies into the encoding–decoding network structure. The comparative quantitative results of different introduction strategies on the omnidirectional datasets are shown in Table 1 (Mask.Strategy). It can be seen from the results of error and accuracy indicators that the introduction of mask map in the encoder on the two datasets is more conducive to obtaining high-accuracy depth estimation result (0.995 for panorama and 0.990 for fisheye in δ_3). This indicates that the mask map plays a more effective role in removing object occlusion in the feature extractor.

4.2.4 Comparison of different depth estimation network

Table 2 presents the comparison results of various depth estimation networks on the omnidirectional datasets, including the widely used depth estimation network FCRN [58] and the network jointing structure recovery and depth estimation, JLDNet [36]. While FCRN is originally designed for depth estimation in traditional perspective images, we modify it to work with omnidirectional datasets and denote it with the prefix “T-”. On the other hand, JLDNet is directly applied for depth estimation of omnidirectional images, and the quantization results show a certain improvement in various indicators. However, it does not consider the problem of distortion and overfitting, resulting in a still high error rate. In contrast, our proposed OmniPDMNet with DAM and USM achieves good performance by improvement in δ_1 from 0.824 to 0.889 for panorama and 0.736 to 0.948 for fisheye, indicating the efficiency of the proposed depth estimation network.

As shown in Fig. 4, the planar depth map estimated by our method exhibits higher consistency with the ground truth compared to other methods, resulting in an overall superior quality of structural depth is higher. This verifies the effectiveness of deepening the network model by introducing up-mapping layers to improve the learning ability. From the perspective of local details, our results have clearer structural boundaries, further proving that the feature-based loss function can better handle the problem of blurred boundaries and significantly improve the prediction accuracy of the planar depth map.

4.3 Performance evaluation of omnidirectional structure recovery

In this section, we first examine the effectiveness of planar depth map for structure recovery on synthetic and real-world datasets. We then provide a qualitative comparison against the state-of-the-art methods on panorama and fisheye datasets and report the numerical comparison and analysis.

4.3.1 Effect of planar depth map on synthetic datasets

To validate the effectiveness of planar depth as geometric prior knowledge for guiding high-quality structure recovery, we conduct comparative experiments before and after introducing the depth map, and the quantitative results are shown in Table 3. Obviously, our OmniSRNet with planar depth map achieves remarkable improvements on synthetic datasets. Especially on fisheye dataset, it exhibits an overall performance gain of 2D IoU (2.1%). Moreover, the visualization results (without or with depth-driven) are shown in Fig. 5. For panorama images ((a)–(c)), the structure recovery results with planar depth-driven are closer to the ground truth structure (marked in blue). Similarly, for fisheye images ((d)–(f)), the structure recovery quality obtained by depth-driven is significantly higher than the result without depth-driven. Among them whether it is the partial occlusion of key corners by small objects or the key corners by larger objects. The depth prior can guide the predicted corners to be more consistent with the annotated corners. The above prove the universality of depth information for various scene structure recovery and provide a guarantee for achieving high-quality 3D point cloud recovery.

4.3.2 Effect of planar depth map on real-world datasets

To thoroughly validate the effectiveness of our OmniSRNet driven by planar depth map on real-world datasets, we conduct extensive experiments and analysis. However, obtaining the planar depth of real scene in the omnidirectional dataset is challenging. To overcome this limitation and perform a comprehensive evaluation, we employ two fine-tuning strategies with different annotations: i) pretraining on synthetic datasets without depth annotation, then fine-tuning on real-world datasets, ii) pre-training on synthetic datasets with depth annotation, then fine-tuning on real-world datasets. The quantitative results are shown in Table 3. Compared to OmniSRNet without depth, with depth-driven exceeds it by 7.1 and 10.5 percent in terms of IoU on panorama and fisheye datasets, respectively. Additionally, the quantitative results (w/ finetune) are superior to those (w/o finetune), indicating that the pretraining model significantly enhances the performance of structure recovery. The qualitative evaluation results are shown in Fig. 6. Obviously, the use of pre-training on synthetic dataset can produce more accurate structure recovery results on real-world datasets ((a)–(c) for panorama images and (d)–(f) for fisheye images). Moreover, the results of pre-training without and with depth map are shown in Fig. 6c and f revealing that pre-training with depth-driven performs better on omnidirectional image and generates more plausible structure recovery.

4.3.3 Comparison with the state-of-the-art structure methods on panorama datasets

We compare our method with previous seven works on omnidirectional datasets, including LayoutNet$_{v1}$ [31], LayoutNet$_{v2}$ [32], CFL$_{std}$ and CFL$_{equi}$ [19], HorizonNet [33], OmniLayout [18], JLDNet [36] and LGTNet [17]. Table 4 presents the comparison results of our method with the state-of-the-art data-driven methods on panorama datasets. Obviously, our OmniSRNet driven by planar depth achieves the best performance on all panorama datasets, including PanoContext, Stanford2D3D and Structured3D. Our OmniSRNet exceeds JLDNet [36] by 2.43, 4.53 and 3.02 percent in terms of 3DIoU on the three datasets, respectively. Specially, OmniSRNet with depth-driven boosts the overall performance on panorama datasets by a margin (3.48%, 3.21% and 2.23% in 3DIoU) compared to LGTNet [17]. This competitive accuracy shows that planar depth as geometric prior knowledge can better alleviate the severe occlusion of complex scenes giving rise to more accurate structure recovery results. Furthermore, we visualize the comparison results of our OmniSRNet to other state-of-the-art methods on panorama datasets (Fig. 7). The first two rows are the comparison results of all methods on real-world panorama scenes. The last four rows demonstrate the comparison results on synthetic panorama scenes. Our OmniSRNet is superior to the other methods on various panorama image and has better robustness in many situations, such as non-cuboid type and open corridors as shown in Fig. 7 (5th-6th rows). The above demonstrates that our method can consistently produce plausible structure recovery results from various scenes even with total or partial object occlusion.

4.3.4 Comparison with the state-of-the-art structure methods on fisheye datasets

To verify the generality and robustness of our OmniSRNet, we further conduct a comparison on fisheye datasets. However, there are relatively few methods related to structure recovery from fisheye image, and presently none of them has released the source code. Therefore, we first modify the source code of LayoutNet$_{v1}$ [31], LayoutNet$_{v2}$ [32], CFL$_{std}$ and CFL$_{equi}$ [19], HorizonNet [33], OmniLayout [18], JLDNet [36] and LGTNet [17] for fisheye image and marked with the prefix 'T-'. Comparisons on three fisheye datasets (PanoContext-F, Stanford2D3D-F and Structured3D-F) are detailed in Table 5. Compared to T-JLDNet, our OmniSRNet with depth-driven presents significant superiority over the competing methods on all fisheye datasets, with an overall performance gain of CE (2.4%, 2.11% and 0.49%), PE (1.66%, 1.21% and 0.49%) and 2D IoU (7.08%, 6.57% and 3.63%), respectively. What's more, OmniSRNet makes a remarkable improvement on these datasets, with a large margin (8.75%, 10.05% and 5.47% in 2DIoU). Figure 8 displays the visual comparison result of our method to the state-of-the-art methods on fisheye datasets. It can be observed that our method consistently provides excellent performance on all of fisheye image, which closely matches the human-annotated ground truth.

5 Applications

This section verifies the effectiveness of the omnidirectional structure recovery from practical applications. It is a useful input for numerous applications, such as video surveillance based on Mixed Reality (MR) and house viewing based on Virtual Reality (VR).

6 MR video surveillance

MR fusion technology [59–65] integrates multiple video streams into 3D space, offers users an immersive visual experience with detailed information and maintains space-time consistency, enhancing the cognitive ability of global spatial information. However, this technology is inefficient and costly in the projection process. Our proposed omnidirectional image 3D structure recovery algorithm presents a promising solution to address this issue. An example of MR-based video surveillance in a real building is shown in Fig. 9.

This scene comes from an office environment in a real video surveillance system of a building. Nine fisheye cameras are deployed in this environment to capture 180° omnidirectional RGB images, as shown in Fig. 9a. Using the omnidirectional image structure recovery network proposed in this paper to predict the 2D structure in the fisheye image (marked by green line segments), as shown in Fig. 9b. The corresponding 3D structure is recovered by the omnidirectional image 3D point cloud method, as shown in Fig. 9c. The recovered 3D structure is registered in the 3D environment, which contains the 3D model corresponding to the office scene. The model is constructed through the artificial CAD model, which is one of the main display elements and provides an overall space for MR video surveillance, as shown in Fig. 9d. Real-time video is projected as texture onto the registered 3D model through texture mapping technology to achieve immersive video surveillance, as shown in Fig. 9e. This technology realizes the simultaneous visualization of multiple video streams and can be used for monitoring and management of smart buildings.

7 VR house viewing

Virtual viewing [66–68] is mainly to replicate the simulated virtual environment of the real or imagined world through virtual reality technology, providing users with a precise and immersive viewing experience in the virtual environment. By leveraging 3D panoramic reality technology, VR house viewing allows users to explore and understand the structure and details of a house online. However, traditional modeling methods for building structures often suffer from high workload, significant cost and lengthy processing cycle. The omnidirectional image 3D structure recovery algorithm proposed in this paper can be used for efficient house structure recovery, improving efficiency and reducing costs. An example is shown in Fig. 10.

For real estate VR viewing, the scene is captured using a panoramic camera, resulting in six 360° omnidirectional RGB images, as shown in Figure (a). Taking the panoramic image as the input, the omnidirectional image structure recovery network proposed in this paper predicts the two-dimensional structure (marked by the green line segment) including the ground, wall and ceiling in the panoramic image, as shown in Fig. 10(b). Utilizing the panoramic projection model, the 3D point cloud structure corresponding to the panoramic image is generated by the point cloud recovery method, as shown in Fig. 10(c). The generated 3D model and 2D floor plan, as shown in Fig. 10(d), ensure that the floor plan and the 3D model can be perfectly matched by data correction, and generate a panoramic display effect of the entire indoor scene for VR viewing, as shown in Fig. 10(e). With a brand-new visualization, contextualization and immersive viewing experience, this technology not only allows users to explore properties in a more context-rich and immersive environment but also fulfills the business requirements for digital exhibitions.

8 Conclusions and future work

In this paper, we propose an effective and efficient approach to generate high-quality structure recovery from a single omnidirectional image. Firstly, we devise a planar depth map learning network (OmniPDMNet), introducing upsampling strategy and adopting a feature-based loss function to improve the accuracy of depth estimation. Then we construct a geometric-driven omnidirectional structure recovery network (OmniSRNet), leveraging the planar depth map as geometric prior to alleviate the key areas interference from cluttered objects and generate high-quality recovery results. We demonstrate the flexibility and effectiveness of OmniSRNet through numerous applications, such as MR-based video surveillance and VR-based house viewing. Finally, a large variety set of experiments are carefully designed and conducted to validate the effectiveness of OmniSRNet on omnidirectional datasets. Experiments demonstrate that our method significantly outperforms the state-of-the-art methods in both quantitative metrics and visual results. To further boost the performance of structure recovery, we will explore a 3D loss-based unified framework without the limitation of Manhattan assumption. We will attempt to extend our OmniSRNet to floor plan recovery from multiple omnidirectional images. What's more, we plan to explore object recovery from omnidirectional image and recover the entire scene layout to enhance perception and understanding of the scene.

Appendix A: Evaluation metrics

A.1: Quantitative metrics of depth estimation

To keep the evaluation metrics for depth estimation consistent with the previous works [21–24, 49], we adopted the following four standard metrics to quantitatively evaluate the performance of our approach. The metrics are:

Absolute Relative Error

(*ABS_REL*), which is the absolute value of the difference between the pixel by pixel predicted depth and ground truth depth. It is normalized by the real depth value, and the normalized sum is normalized by the total number of pixels, and defined as

$$ABS\_REL = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\| d_i^{pd} - g_i^{gt} \right\|}{g_i^{gt}} \tag{A1}$$

where N is the pixel number of ground truth. d_i^{pd} and g_i^{gt} denote the depth value of the predicted and ground truth, respectively. The lower of this metric, the accuracy of the network model, and the better the result of depth estimation.

Square Relative Error (*SQ_REL*), which is the absolute value square of the difference between the pixel by pixel predicted depth and ground truth depth. It is also normalized by the real depth value, and the normalized sum is normalized by the total number of pixels, and defined as

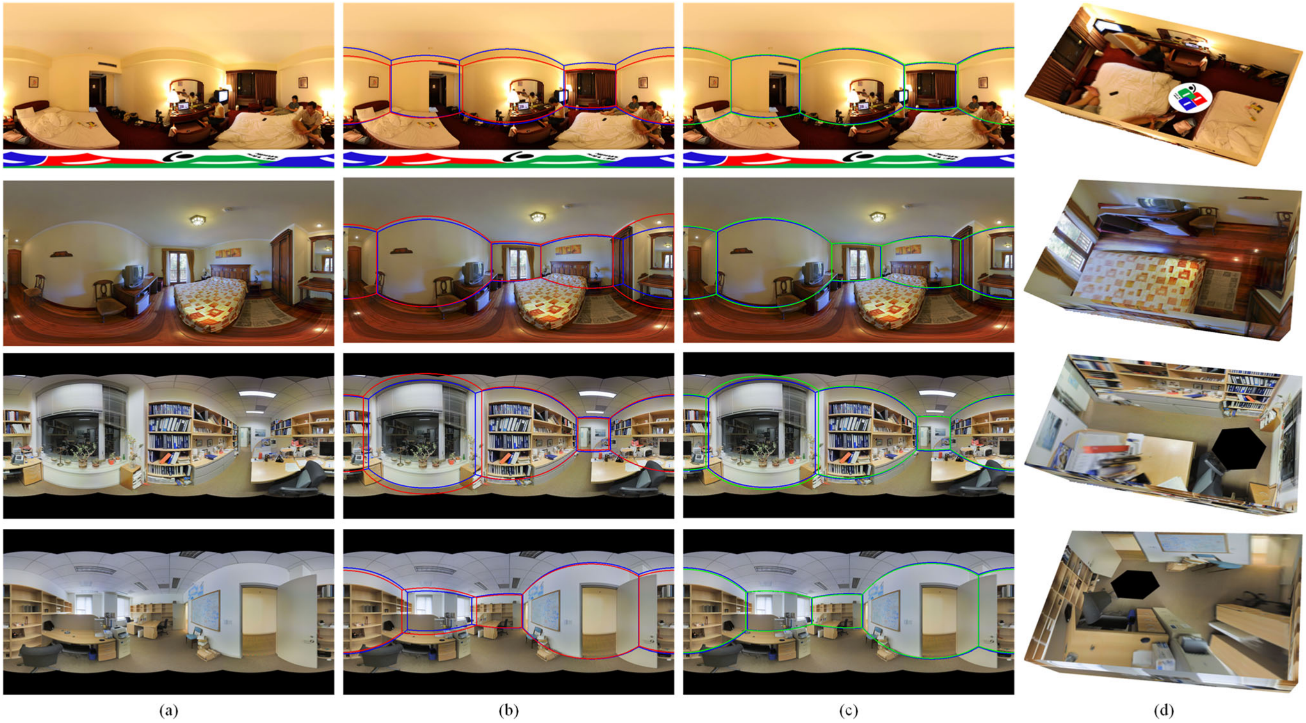

Fig. 11 Qualitative comparison of the effect of planar depth map for structure recovery on panorama real dataset. Left to Right: For each panorama image, we show its original image (**a**), and the structure recovered by our method w/o depth-driven (**b**), w/ depth-driven (**c**), 3D point cloud (**d**). The predictions of our method with and without depth-driven are highlighted in green and red, respectively, whereas the ground truth is in blue

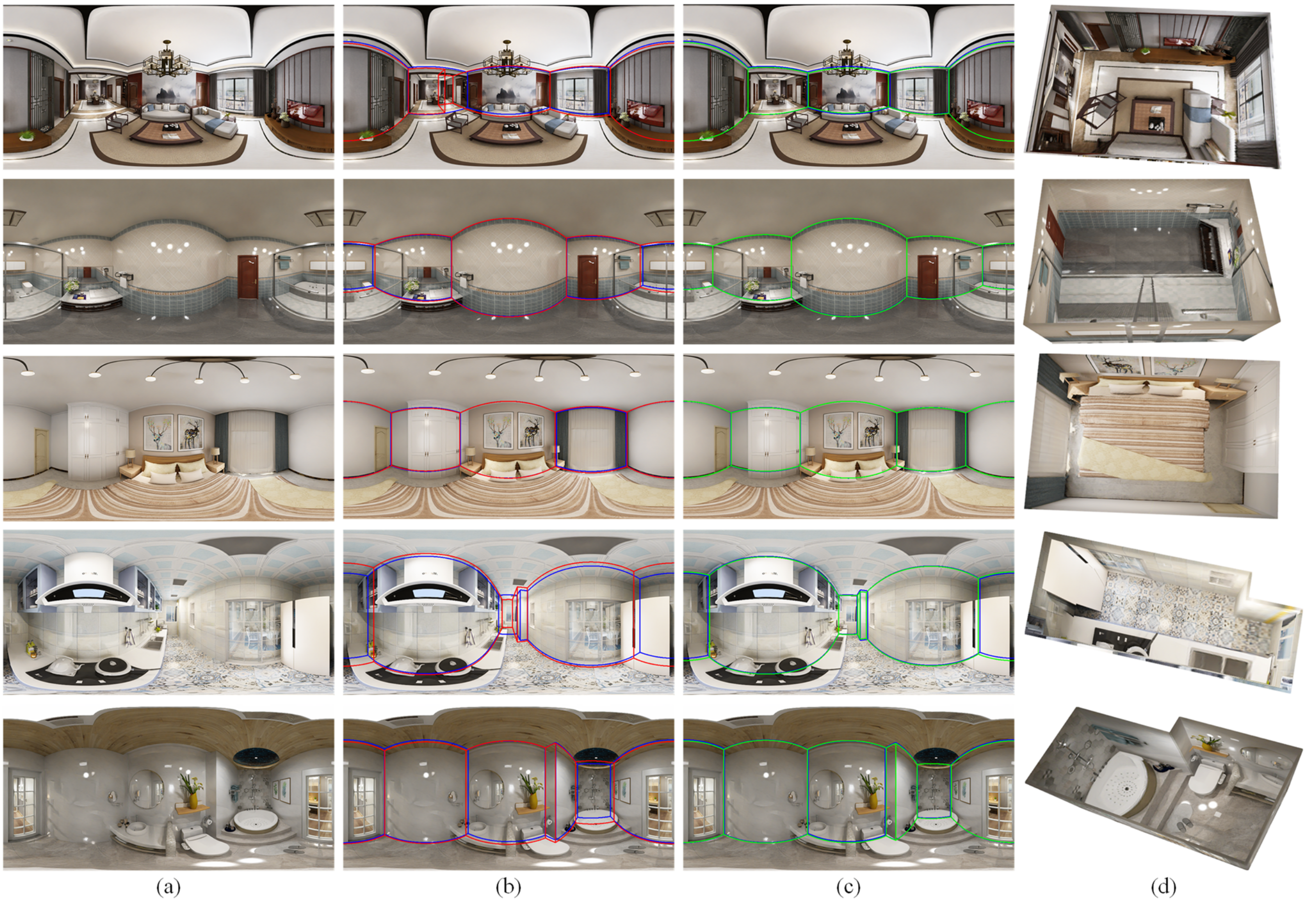

Fig. 12 Qualitative comparison of the effect of planar depth map for structure recovery on panorama synthetic dataset. Left to Right: For each panorama image, we show its original image (**a**), and the structure recovered by our method w/o depth-driven (**b**), w/ depth-driven (**c**), 3D point cloud (**d**). The predictions of our method with and without depth-driven are highlighted in green and red, respectively, whereas the ground truth is in blue

$$SQ\_REL = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|d_i^{pd} - g_i^{gt}\right\|^2}{g_i^{gt}} \tag{A2}$$

where N is the pixel number of ground truth. d_i^{pd} and g_i^{gt} denote the depth value of the predicted and ground truth, respectively. The lower of this metric, the accuracy of the network model, and the better the result of depth estimation.

Root Mean Square Error (*RMSE*), which represents the depth difference between the predicted structure depth and ground truth, and defined as

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{i=1}^{N}\left\|d_i^{pd} - g_i^{gt}\right\|^2} \tag{A3}$$

where N is the pixel number of ground truth. This metric is mainly used to evaluate the accuracy of non-cuboid 3D structure recovery, and the lower the value, the better.

Percentage of Pixels (δ), which is defined as the percentage of pixels with the ratio (or its reciprocal) between predicted depth and ground truth depth smaller than the thread T, as follows

$$\max\left(\frac{d^{pd}}{d^{gt}}, \frac{d^{gt}}{d^{pd}}\right) = \delta < \mathcal{T} \tag{A4}$$

where $T = 1.25$. The higher the value of this metric, the better.

A.2: Quantitative metrics of structure recovery

To keep the evaluation metrics for structure recovery consistent with the previous works [17–19, 31, 33, 36], we adopted the following four standard metrics to quantitatively evaluate the performance of our approach. The metrics are:

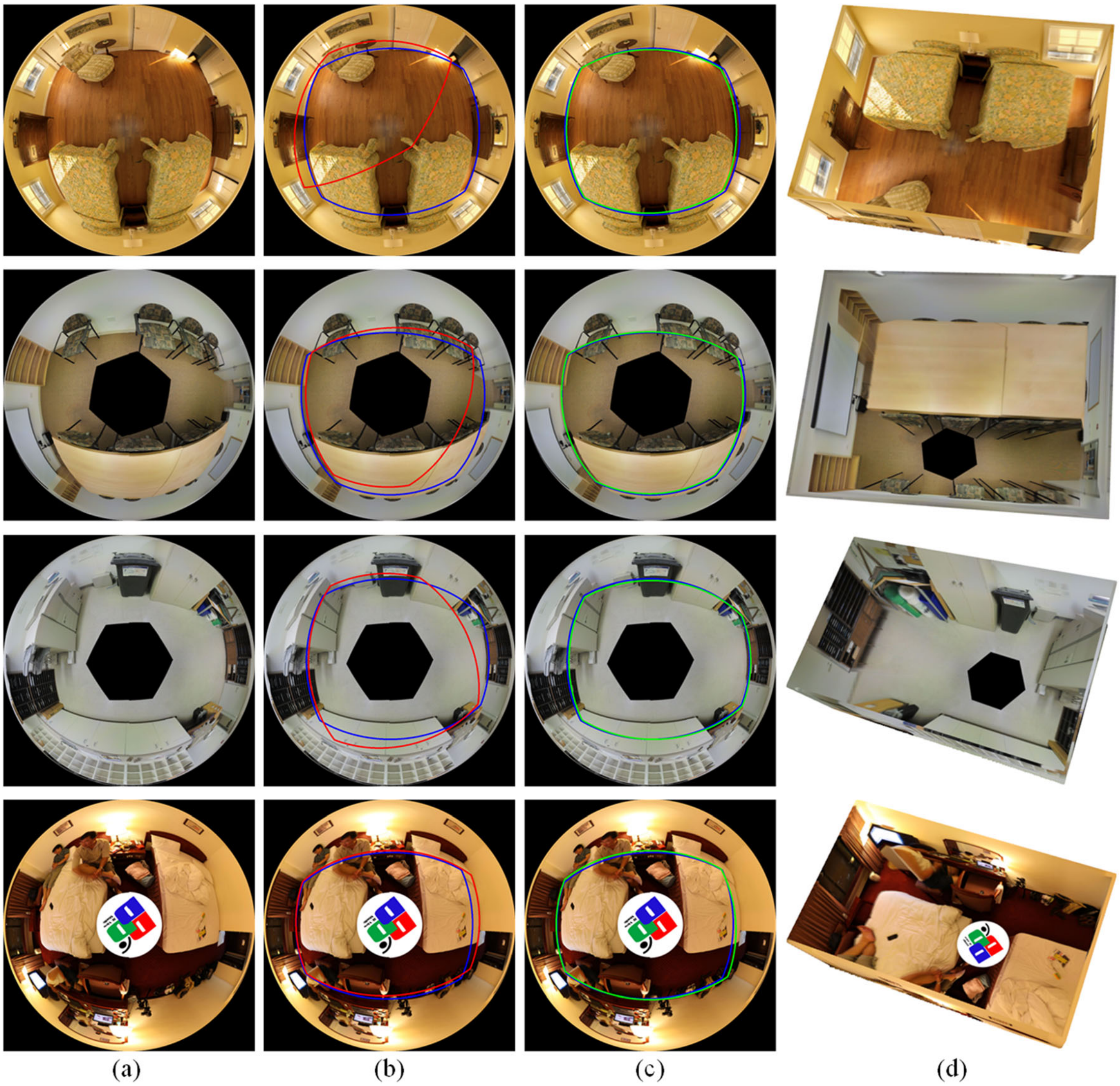

Fig. 13 Qualitative comparison of the effect of planar depth map for structure recovery on fisheye real dataset. Left to Right: For each fisheye image, we show its original image (**a**), and the structure recovered by our method w/o depth-driven (**b**), w/ depth-driven (**c**), 3D point cloud (**d**). The predictions of our method with and without depth-driven are highlighted in green and red, respectively, whereas the ground truth is in blue

Corner Error(*CE*), which is the normalized *L*2 distance between predicted corners and ground truth corners across all images, and defined as

$$CE = \frac{\sum_{i=1}^{N_c} \left\| c_i^{pd} - c_i^{gt} \right\|_2^2}{\sqrt{H^2 + W^2}} \tag{A5}$$

where N_c is the number of corners in structure. H and W are the height and width of the image. For panorama image $H = 1024$, $W = 512$, while for fisheye image $H = W = 512$. c_i^{pd} and c_i^{gt} denote the position coordinates of the predicted and ground truth corners, respectively. The lower of this metric, the accuracy of the network model, and the better the result of structure recovery.

Pixel Error (*PE*), which is the pixel-wise error between the predicted plane classes (ceil, wall and floor for panorama; wall and floor for fisheye) of structure and the ground truth across all images, and defined as

$$PE = \frac{\sum_{i=1}^{N_p} \mathcal{H}(p_i \neq g_i)}{W \cdot H} \tag{A6}$$

where N_p is the number of pixels in structure. p_i and g_i denote the pixel value of the predicted and ground truth, respectively. $\mathcal{H}(\cdot)$ is an indicator function, with $\mathcal{H}(\cdot) = 1$ if

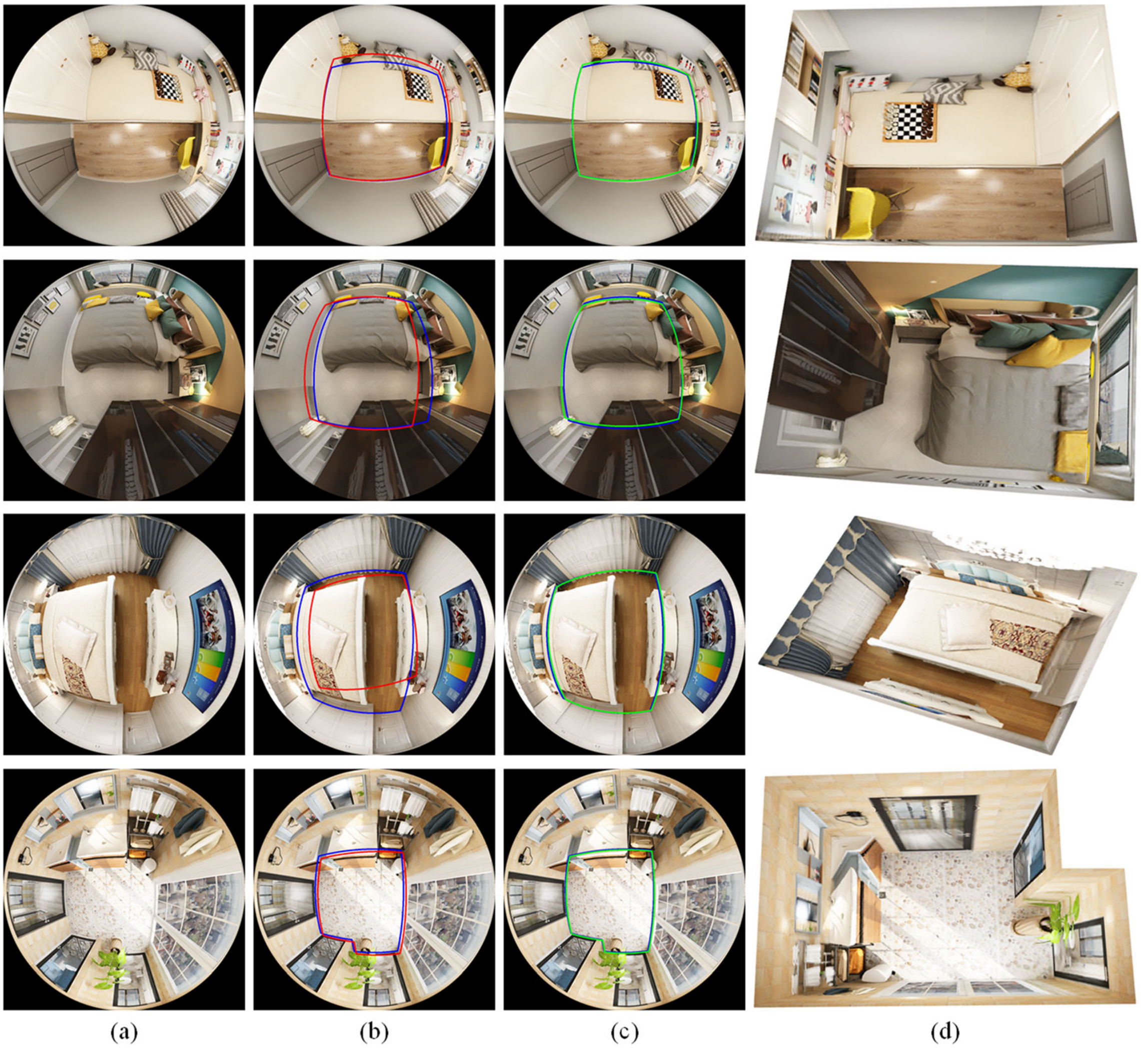

Fig. 14 Qualitative comparison of the effect of planar depth map for structure recovery on fisheye synthetic dataset. Left to Right: For each fisheye image, we show its original image (**a**), and the structure recovered by our method w/o depth-driven (**b**), w/ depth-driven (**c**), 3D point cloud (**d**). The predictions of our method with and without depth-driven are highlighted in green and red, respectively, whereas the ground truth is in blue

$p_i = g_i$ and 0 otherwise. This metric can evaluate the global structure, and the lower the value, the better.

2D Intersection over Union (2*DIoU*), which calculates the pixel-wise intersection-over-union between predicted 2D structure under ceiling view and ground truth for fisheye image, and defined as

$$IoU = \frac{V_2^{pd} \cap V_2^{gt}}{V_2^{pd} \cup V_2^{gt}} \tag{A7}$$

where V_2^{pd} and V_2^{gt} stand for the floor plane occupancy of the predicted and ground truth, respectively. This metric can evaluate 2D accuracy for global structure, and the higher the value, the better.

3D Intersection over Union (3*DIoU*), which calculates the pixel-wise intersection-over-union between predicted 3D structure and ground truth for panorama image, and defined as

$$IoU = \frac{V_3^{pd} \cap V_3^{gt}}{V_3^{pd} \cup V_3^{gt}} \tag{A8}$$

where V_3^{pd} and V_3^{gt} represent the 3D structure occupancy of the predicted and ground truth, respectively. This metric can evaluate 3D accuracy for global structure, and the higher the value, the better.

Appendix B: More results of our method

B.1: Effectiveness of planar depth map for structure recovery

We provide more structure recovery results to validate the effectiveness of planar depth map on the omnidirectional datasets, containing panorama and fisheye dataset. The structure recovery results from panorama image are shown in Figs. 11, 12, 13 and 14 display the structure recovery results from fisheye image. For each result, we show its original image, the structure recovered by our method w/o and w/ depth-driven and 3D point cloud, respectively.

B.2: Comparisons results with structure recovery methods

We report the comparison results with state-of-the-art structure recovery methods on omnidirectional datasets, containing panorama and fisheye dataset. For each panorama image, we present the comparison results with HorizonNet [33], JLDNet [36] and LGTNet [17], and they are shown in Figs. 15 and 16. Additionally, since there is currently no public available code for fisheye structure recovery, we firstly modify the author-provided code of

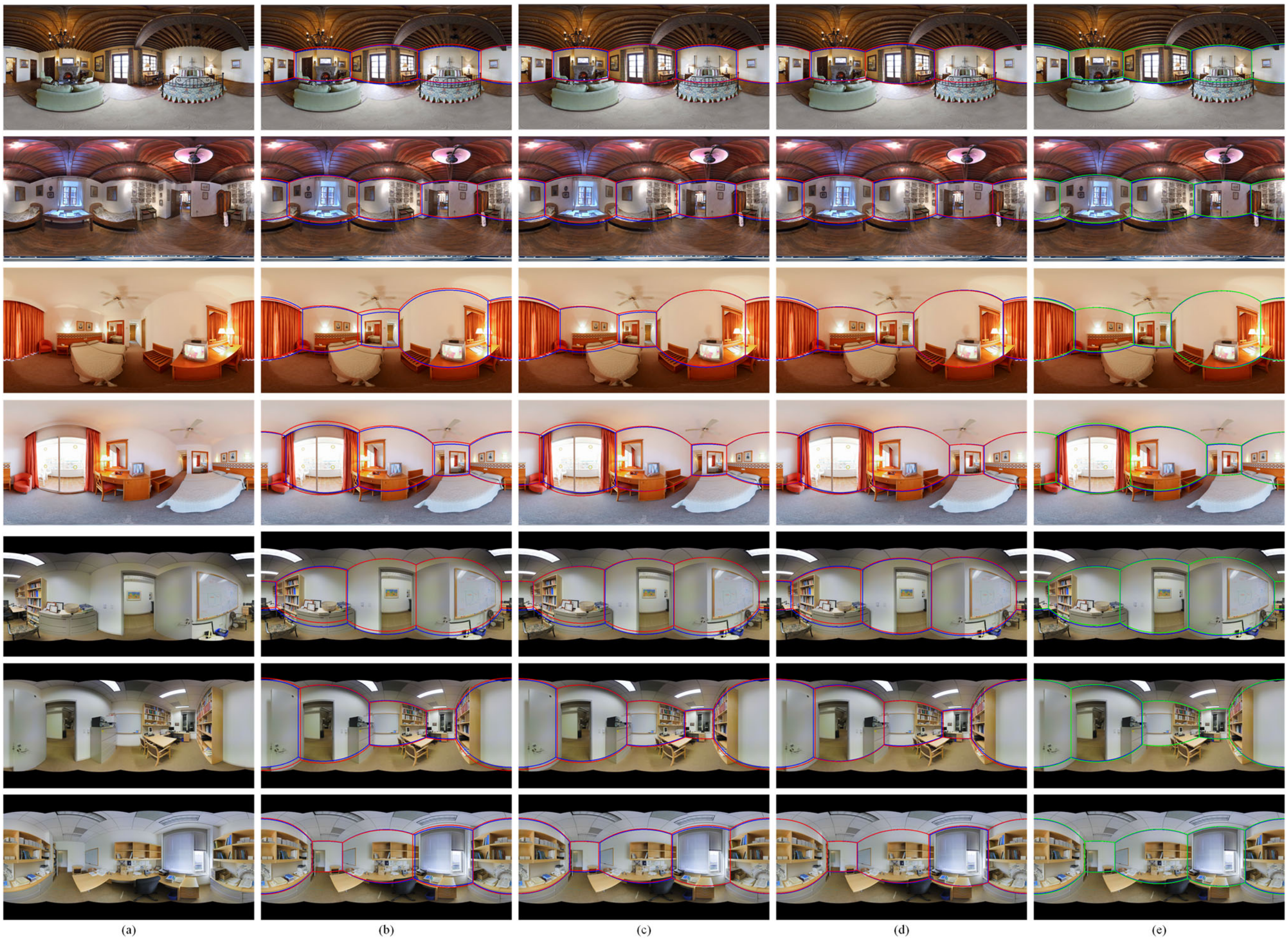

Fig. 15 Qualitative comparison of different methods on panorama real dataset. Left to Right: For each panorama image, we show its original image (**a**), and the structure recovered by HorizonNet (**b**), JLDNet (**c**), LGTNet (**d**), our method with depth-driven (**e**). The predictions of our method and others driven are highlighted in green and red, respectively, whereas the ground truth is in blue

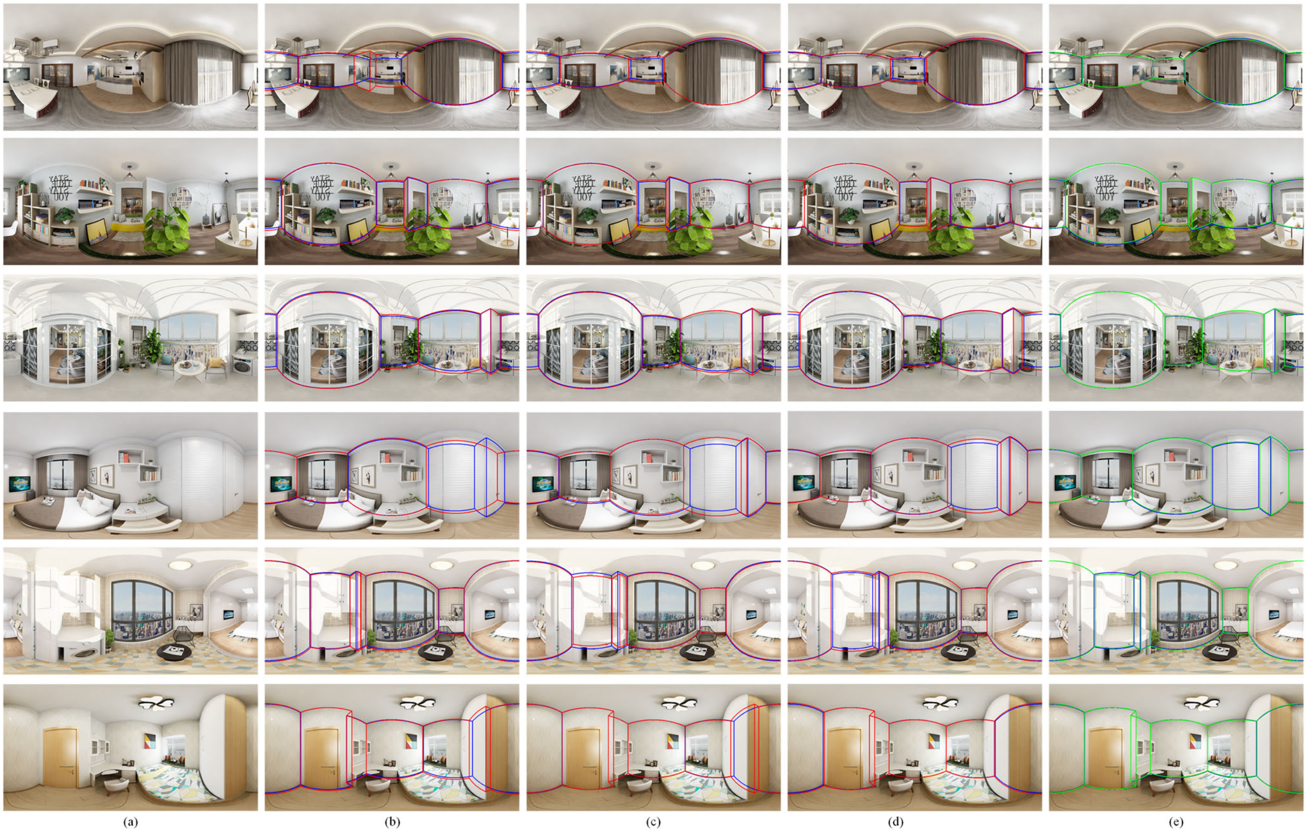

Fig. 16 Qualitative comparison of different methods on panorama synthetic dataset. Left to Right: For each panorama image, we show its original image (**a**), and the structure recovered by HorizonNet (**b**), JLDNet (**c**), LGTNet (**d**), our method with depth-driven (**e**). The predictions of our method and others driven are highlighted in green and red, respectively, whereas the ground truth is in blue

these methods to work with fisheye image and marked with the prefix "T-", such as T-HorizonNet, T-JLDNet and T-LGTNet. And the comparison results are shown Figs. 17 and 18.

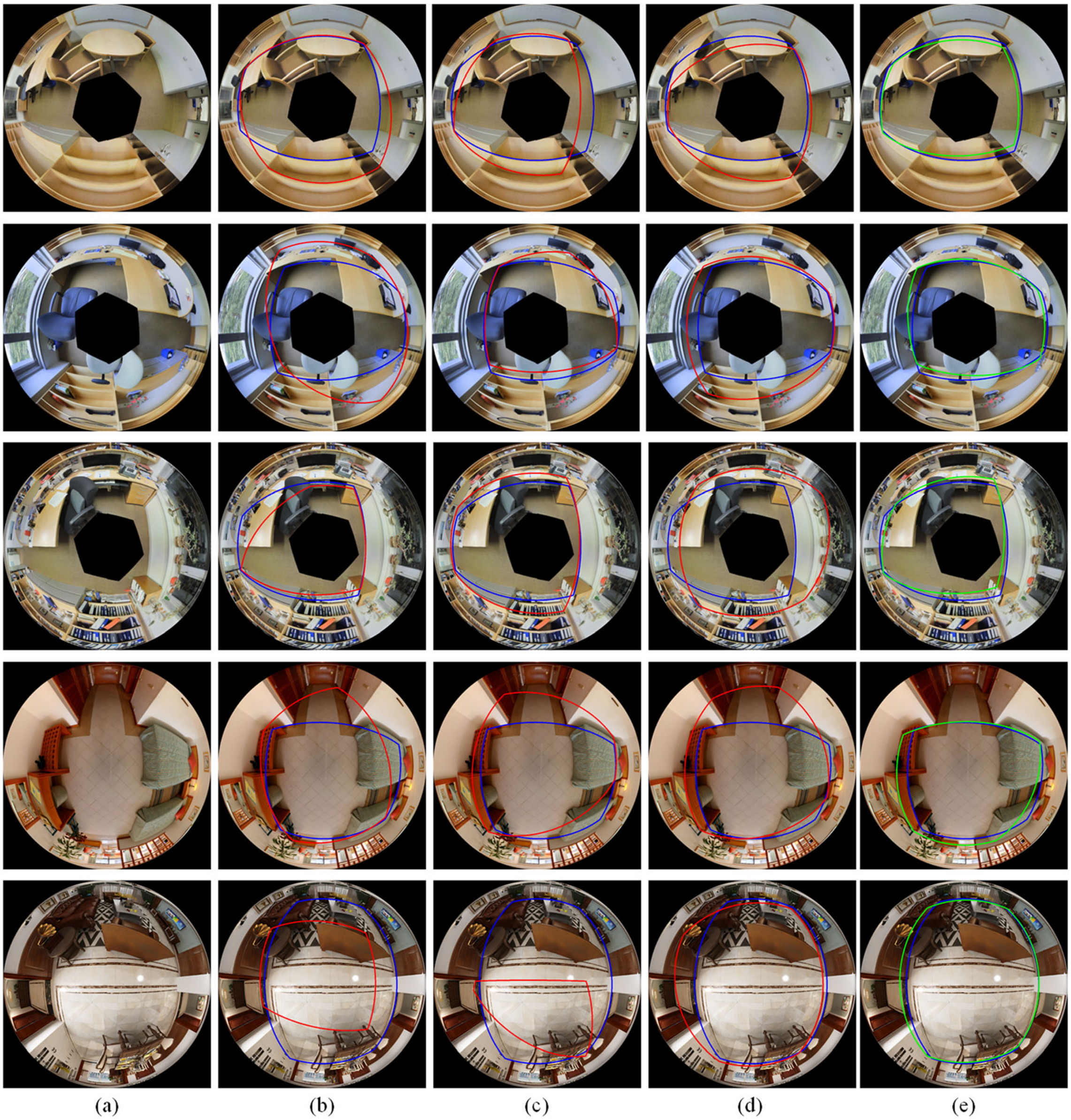

Fig. 17 Qualitative comparison of different methods on fisheye real dataset. Left to Right: For each fisheye image, we show its original image (**a**), and the structure recovered by T-HorizonNet (**b**), T-JLDNet (**c**), T-LGTNet (**d**), our method with depth-driven (**e**). The predictions of our method and others driven are highlighted in green and red, respectively, whereas the ground truth is in blue

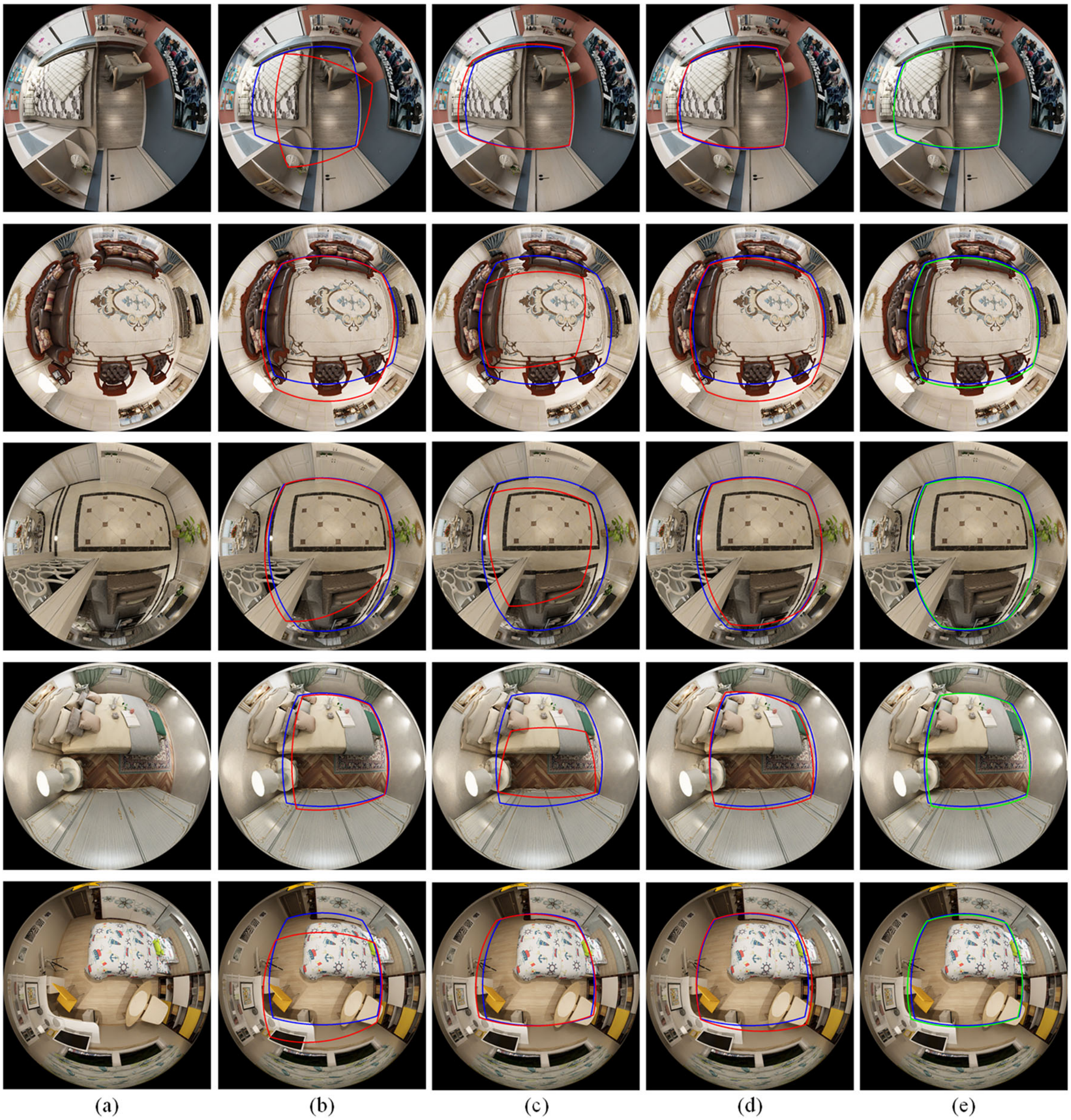

Fig. 18 Qualitative comparison of different methods on fisheye synthetic dataset. Left to Right: For each fisheye image, we show its original image (**a**), and the structure recovered by T-HorizonNet (**b**), T-JLDNet (**c**), T-LGTNet (**d**), our method with depth-driven (**e**). The predictions of our method and others driven are highlighted in green and red, respectively, whereas the ground truth is in blue

Data availability Data available on request from the authors.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Su Y-C, Grauman K (2017) In: 2017 IEEE Conference on Computer Vision And title=Making 360° Video Watchable in 2D: Learning Videography for Click Free Viewing, Pattern Recognition (CVPR), pp 1368–1376
2. Ramakrishnan SK, Al-Halah Z, Grauman K (2020) Occupancy anticipation for efficient exploration and navigation. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds.) Proceedings of the European Conference on Computer Vision (ECCV), pp 400–418

3. Saito H, Baba S, Kanade T (2003) Appearance-based virtual view generation from multicamera videos captured in the 3d room. IEEE Trans Multimedia 5(3):303–316
4. Albanis G, Gkitsas V, Zioulis N, Onsori-Wechtitsch S, Whitehand R, Ström P, Zarpalas D (2023) An ai-based system offering automatic dr-enhanced ar for indoor scenes. In: Nakamatsu K, Patnaik S, Kountchev R, Li R, Aharari A (eds.) Advanced Intelligent Virtual Reality Technologies, pp 187–199
5. Sankar A, Seitz SM (2017) Interactive room capture on 3d-aware mobile devices. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp 415–426
6. Da Silveira TLT, Jung CR (2022) Visual computing in 360°: Foundations, challenges, and applications. In: 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), vol 1, pp 302–307
7. Zhang C, Cui Z, Chen C, Liu S, Zeng B, Bao H, Zhang Y (2021) Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 12632–12641
8. Gkioxari G, Ravi N, Johnson J (2022) Learning 3d object shape and layout without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1695–1704
9. Jia H, Yi H, Fujiki H, Zhang H, Wang W, Odamaki M (2022) 3d room layout recovery generalizing across manhattan and non-manhattan worlds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5192–5201
10. Hedau V, Hoiem D, Forsyth D (2009) Recovering the spatial layout of cluttered rooms. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV '09)
11. Wang H, Hutchcroft W, Li Y, Wan Z, Boyadzhiev I, Tian Y, Kang SB (2022) Psmnet: Position-aware stereo merging network for room layout estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8616–8625
12. Zhang Y, Song S, Tan P, Xiao J (2014) Panocontext: A whole-room 3d context model for panoramic scene understanding. In: European Conference on Computer Vision, pp 668–686
13. Yang H, Zhang H (2016) Efficient 3d room shape recovery from a single panorama. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5422–5430
14. Yang Y, Jin S, Liu R, Kang SB, Yu J (2018) Automatic 3d indoor scene modeling from single panorama. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3926–3934
15. Fernandez-Labrador C, Perez-Yus A, Lopez-Nicolas G, Guerrero JJ (2018) Layouts from panoramic images with geometry and deep learning. In: IEEE Robotics and Automation Letters, vol 3, pp 3153–3160
16. Li M, Zhou Y, Meng M, Wang Y, Zhou Z (2019) 3d room reconstruction from a single fisheye image. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp 1–8
17. Jiang Z, Xiang Z, Xu J, Zhao M (2022) Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1644–1653
18. Rao S, Kumar V, Kifer D, Giles CL, Mali A (2021) Omnilayout: Room layout reconstruction from indoor spherical panoramas. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 3706–3715
19. Fernandez-Labrador C, Facil JM, Perez-Yus A, Demonceaux C, Civera J, Guerrero JJ (2020) Corners for layout: End-to-end layout recovery from 360 images. In: IEEE Robotics and Automation Letters, vol 5, pp 1255–1262
20. Ruder M, Dosovitskiy A, Brox T (2018) Artistic style transfer for videos and spherical images. Int J Comput Vision 126(11):1199–1219
21. Wang F-E, Yeh Y-H, Sun M, Chiu W-C, Tsai Y-H (2020) Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 459–468
22. Jiang H, Sheng Z, Zhu S, Dong Z, Huang R (2021) Unifuse: unidirectional fusion for 360 panorama depth estimation. IEEE Robot Autom Lett 5:1–1
23. Cheng X, Wang P, Zhou Y, Guan C, Yang R (2020) Omnidirectional depth extension networks. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp 589–595
24. Chen H-X, Li K, Fu Z, Liu M, Chen Z, Guo Y (2021) Distortion-aware monocular depth estimation for omnidirectional images. IEEE Signal Process Lett 5:334–338
25. Coughlan JM, Yuille AL (2000) The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA
26. Schwing A, Hazan T, Pollefeys M, Urtasun R (2012) Efficient structured prediction for 3d indoor scene understanding. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, pp 2815–2822
27. Hedau V, Hoiem D, Forsyth D (2010) Thinking inside the box: Using appearance models and context based on room geometry. In: European Conference on Computer Vision
28. Pero LD, Bowdish J, Kermgard B, Hartley E, Barnard K (2013) Understanding bayesian rooms using composite 3d object models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 153–160
29. Xu J, Stenger B, Kerola T, Tung T (2017) Pano2cad: Room layout from a single panorama image. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 354–362
30. Yang S-T, Wang F-E, Peng C-H, Wonka P, Sun M, Chu H-K (2019) Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3358–3367
31. Zou C, Colburn A, Shan Q, Hoiem D (2018) Layoutnet: reconstructing the 3d room layout from a single rgb image. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2051–2059
32. Zou C, Su JW, Peng CH, Colburn A, Shan Q, Wonka P, Chu HK, Hoiem D (2021) Manhattan room layout reconstruction from a single 360° image: a comparative study of state-of-the-art methods. International Journal of Computer Vision, pp 1–22
33. Sun C, Hsiao C-W, Sun M, Chen H-T (2019) Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1047–1056
34. Pérez-Yus A, López-Nicolás G, Guerrero JJ (2016) Peripheral expansion of depth information via layout estimation with fisheye camera. In: European Conference on Computer Vision, pp 396–412
35. Zhang W, Zhang W, Zhang Y (2020) Geolayout: Geometry driven room layout estimation based on depth maps of planes. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 632–648
36. Zeng W, Karaoglu S, Gevers T (2020) Joint 3d layout and depth prediction from a single indoor panorama image. In: 16th European Conference, Glasgow, UK, August 23-28, 2020, pp 666–682

37. Dong X, Garratt MA, Anavatti SG, Abbass HA (2022) Towards real-time monocular depth estimation for robotics: a survey. IEEE Trans Intell Transp Syst 23(10):16940–16961
38. Sayed M, Gibson J, Watson J, Prisacariu V, Firman M, Godard C (2022) Simplerecon: 3d reconstruction without 3d convolutions. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds.) Proceedings of the European Conference on Computer Vision (ECCV), pp 1–19. Springer
39. Gao S, Yang K, Shi H, Wang K, Bai J (2022) Review on panoramic imaging and its applications in scene understanding. IEEE Trans Instrum Meas 71:1–34
40. Hoiem D, Efros AA, Hebert M (2005) Geometric context from a single image. In: Tenth IEEE International Conference on Computer Vision, pp 654–661
41. Liu B, Gould S, Koller D (2010) Single image depth estimation from predicted semantic labels. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 1253–1260
42. Han C, Cheng D, Kou Q, Wang X, Chen L, Zhao J (2022) Self-supervised monocular depth estimation with multi-scale structure similarity loss. Multimedia Tools Appl 6:1–16
43. Xu Q, Kong W, Tao W, Pollefeys M (2022) Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. In: IEEE Transactions on Pattern Analysis and Machine Intelligence
44. Zhou Z, Dong Q (2022) Self-distilled feature aggregation for self-supervised monocular depth estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 709–726
45. Zhuang C, Lu Z, Wang Y, Xiao J, Wang Y (2022) Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 36, pp 3653–3661
46. Tateno K, Navab N, Tombari F (2018) Distortion-aware convolutional filters for dense prediction in panoramic images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 732–750
47. Zioulis N, Karakottas A, Zarpalas D, Daras P (2018) Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 453–471
48. Eder M, Moulon P, Guan L (2019) Pano popups: Indoor 3d reconstruction with a plane-aware network. In: 2019 International Conference on 3D Vision (3DV), pp 76–84
49. Jin L, Xu Y, Zheng J, Zhang J, Tang R, Xu S, Yu J, Gao S (2020) Geometric structure based and regularized depth estimation from 360 indoor imagery. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 886–895
50. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778
51. Nie Y, Han X, Guo S, Zheng Y, Chang J, Zhang JJ (2020) Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 52–61
52. Meng M, Xiao L, Zhou Y, Li Z, Zhou Z (2021) Distortion-aware room layout estimation from a single fisheye image. In: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp 441–449
53. Armeni I, Sax S, Zamir AR, Savarese S (2017) Joint 2D-3D-semantic data for indoor scene understanding
54. Zheng J, Zhang J, Li J, Tang R, Gao S, Zhou Z (2019) Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 519–535
55. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 764–773
56. Zhu X, Hu H, Lin S, Dai J (2019) Deformable convnets v2: more deformable, better results. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9308–9316
57. Issaranon T, Zou C, Forsyth D (2019) Counterfactual depth from a single rgb image. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp 2129–2138
58. Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N (2016) Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp 239–248
59. Cui X, Khan D, He Z, Cheng Z (2023) Fusing surveillance videos and three-dimensional scene: a mixed reality system. Comput Anim Virtual Worlds 34(1):1–15
60. Büschel W, Lehmann A, Dachselt R (2021) Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp 1–15
61. Büschel W, Lehmann A, Dachselt R (2021) Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp 1–15
62. Philip D, George S, Rony K, Deb R (2010) An immersive system for browsing and visualizing surveillance video. In: ACM International Conference on Multimedia, pp 371–380
63. Zhou Y, Cao M, You J, Meng M, Wang Y, Zhou Z (2018) MR video fusion: interactive 3D modeling and stitching on wide-baseline videos. In: ACM Symposium on Virtual Reality Software and Technology, p 17
64. Zhou Z, Meng M, Zhou Y, Zhu Z, You J (2021) Model-guided 3d stitching for augmented virtual environment. Sci China Inf Sci 5:96
65. Zhu G, Zhang H, Jiang Y, Lei J, He L, Li H (2023) Dynamic fusion technology of mobile video and 3d gis: the example of smartphone video. ISPRS Int J Geo Inf 12(3):125
66. Azmi A, Ibrahim R, Abdul Ghafar M, Rashidi A (2022) Smarter real estate marketing using virtual reality to influence potential homebuyers' emotions and purchase intention. Smart Sustain Built Environ 11(4):870–890
67. Chhikara P, Kuhar H, Goyal A, Sharma C (2023) Digitour: Automatic digital tours for real-estate properties. In: Proceedings of the 6th Joint International Conference on Data Science and Management of Data, pp 223–227
68. Mendes NP, Santos ET (2022) Exploratory virtual model: study and evaluation of a low-cost vr-based real estate sales tool. J Geom Gr 26(1):171–184

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.