Monocular Occupancy Prediction for Scalable Indoor Scenes

Hongxiao Yu[1,2], Yuqi Wang[1,2], Yuntao Chen[3], and Zhaoxiang Zhang[1,2,3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[2] NLPR, MAIS, Institute of Automation, Chinese Academy of Sciences (CASIA)
[3] Centre for Artificial Intelligence and Robotics (HKISI_CAS)
{yuhongxiao2023, wangyuqi2020, zhaoxiang.zhang}@ia.ac.cn,
chenyuntao08@gmail.com

Abstract. Camera-based 3D occupancy prediction has recently garnered increasing attention in outdoor driving scenes. However, research in indoor scenes remains relatively unexplored. The core differences in indoor scenes lie in the complexity of scene scale and the variance in object size. In this paper, we propose a novel method, named ISO, for predicting indoor scene occupancy using monocular images. ISO harnesses the advantages of a pretrained depth model to achieve accurate depth predictions. Furthermore, we introduce the Dual Feature Line of Sight Projection (D-FLoSP) module within ISO, which enhances the learning of 3D voxel features. To foster further research in this domain, we introduce *Occ-ScanNet*, a large-scale occupancy benchmark for indoor scenes. With a dataset size 40 times larger than the NYUv2 dataset, it facilitates future scalable research in indoor scene analysis. Experimental results on both NYUv2 and Occ-ScanNet demonstrate that our method achieves state-of-the-art performance. The dataset and code are made publicly at `https://github.com/hongxiaoy/ISO.git`.

Keywords: 3D occupancy prediction · Camera-based scene understanding · Semantic scene completion

1 Introduction

3D scene understanding is a crucial task in computer vision, becoming increasingly important for applications such as robotic navigation [10], augmented reality [2], and autonomous driving [47]. While humans possess a natural ability to comprehend 3D environments through vision, enabling this ability in computers poses a significant challenge due to limitations like restricted fields of view, sparse sensing capabilities, and measurement noise. However, with the rapid development of deep learning and the availability of large-scale 3D driving datasets, camera-based 3D object detection [14, 15, 20, 21, 35, 37, 39, 43] and occupancy prediction [5, 22, 33, 38, 40] have developed rapidly, achieving significant improvements in performance.

Occupancy prediction has gained popularity recently due to its effectiveness in representing both background and foreground objects within a scene using a unified representation. Although significant progress has been made in outdoor driving scenarios, research in indoor scenarios remains limited. Indoor scenes differ from outdoor driving scenes in two key aspects: (1) *scene-scale complexity*: Indoor rooms often exhibit a more diverse range of sizes compared to outdoor environments, where driving scenarios typically focus on a fixed 3D space for perception. This diversity, ranging from spacious living rooms to narrow kitchens, poses higher precision requirements for depth prediction. (2) *object complexity*: Indoor scenes feature a higher density and greater variety of objects. Unlike outdoor objects such as vehicles and pedestrians, which typically have consistent sizes within their respective categories and are relatively well-separated, indoor furniture and other objects often exhibit significant variations in scale and are closely positioned to each other. This increased complexity necessitates more sophisticated 3D perception techniques to accurately capture and understand the intricate geometry and relationships among objects within indoor spaces. Furthermore, existing works [5,45] focusing on indoor scenes primarily utilize the NYUv2 dataset [29], which lacks a more scalable and generalizable benchmark for comprehensive evaluation.

To address the above issues, we firstly introduce a more scalable benchmark, *Occ-ScanNet*, for 3D occupancy prediction in indoor scenes. This benchmark builds upon the large-scale ScanNet [8] dataset, offering a 40 times more samples compared to the NYUv2 [29] dataset, thus greatly expanding the scope and diversity of indoor scene study. To effectively address the unique complexities of indoor scenes, we propose a novel method named ISO (Indoor Scene Occupancy). ISO leverages a powerful pretrained depth model and integrates a D-FLoSP (Dual Feature Line of Sight Projection) module. This module enables precise depth estimation and facilitates learning voxel features for accurate predictions. To accommodate the varying sizes of indoor objects, we introduce a multi-scale feature fusion module. This module enhances the learning of object features across different scales. Leveraging a powerful depth model, our method is adept at handling diverse indoor scenes, offering a robust and versatile solution.

Our main contributions can be summarized as follows.

- We introduce a new benchmark called *Occ-ScanNet* for monocular 3D occupancy prediction in indoor scenes. With a dataset size 40 times larger than the NYUv2 dataset, it significantly enhances the potential for scalable research in indoor scene analysis.
- We propose a novel approach called *ISO*, which primarily comprises the D-FLoSP (Dual Feature Line of Sight Projection) module and a multi-scale feature fusion module. Together, these components effectively address the challenges posed by variations in scene and object sizes, enabling more accurate and robust 3D occupancy prediction.
- Experiments on the Occ-ScanNet and NYUv2 datasets demonstrate that our approach achieves state-of-the-art performance in monocular 3D occupancy prediction. Furthermore, our method exhibits significant scalability potential.

2 Related works

2.1 Monocular 3D Semantic Scene Completion

Monocular 3D Semantic Scene Completion (SSC) aims to infer the complete 3D structure and corresponding semantics from a single image. This monocular setting was first introduced by Monoscene [5], advancing upon prior SSC methods [18, 23, 30, 48, 51] by relying solely on vision, without additional 3D inputs. Monoscene [5] introduces the Features Line of Sight Projection (FLoSP), inspired by optics, to obtain 3D features through ray projection. However, the shared 2D features lifted to 3D rays via FLoSP results in depth ambiguity. Consequently, subsequent efforts have placed greater emphasis on leveraging depth information. VoxFormer [19] proposed a novel query proposal network based on 2D convolutions, generating sparse queries from image depth, which showed impressive performance in driving scenes. Meanwhile, NDC-Scene [45] further devised a Depth-Adaptive Dual Decoder to concurrently upsample and merge the 2D and 3D feature maps, thereby enhancing overall performance.

2.2 Multiview 3D Occupancy Prediction

Multiview 3D Occupancy Prediction predicts the semantic occupancy of the surrounding 3D scene given multiview images, has recently garnered significant attention in the field of autonomous driving. TPVFormer [17] pioneers exploration in driving scenes, employing sparse LiDAR labels for supervision. It introduces a tri-perspective view (TPV) representation that accompanies BEV with two additional perpendicular planes. However, due to the sparse supervision provided by LiDAR, subsequent works [33, 34, 36, 40] have focused on providing more dense occupancy benchmarks. FB-OCC [22] integrates the lifting of 2D to 3D and querying of 3D to 2D, achieving a more efficient 3D feature transformation. Meanwhile, PanoOcc [38] unifies detection and semantic occupancy tasks, enabling comprehensive panoramic scene understanding. Recent methods also explore improving model efficiency [46] and utilizing weaker forms of supervision [16, 25].

2.3 3D Reconstruction from Image

3D reconstruction is a technique that involves recovering a 3D representation of objects [1, 11, 26], scenes [9, 41, 49], or even human bodies [12, 13, 50] from camera images. The task of 3D reconstruction can be classified into two categories: monocular reconstruction and multi-view reconstruction, depending on the number of images utilized. In the context of indoor scenes, 3D scene reconstruction aims to determine the surface geometry of the entire scene, often without incorporating semantic information. Atlas [24] proposes an end-to-end 3D reconstruction framework using TSDF regression from RGB images, bypassing traditional depth map estimation for efficient semantic segmentation. Panoptic-Reconstruction [7] unifies geometric reconstruction, 3D semantic, and

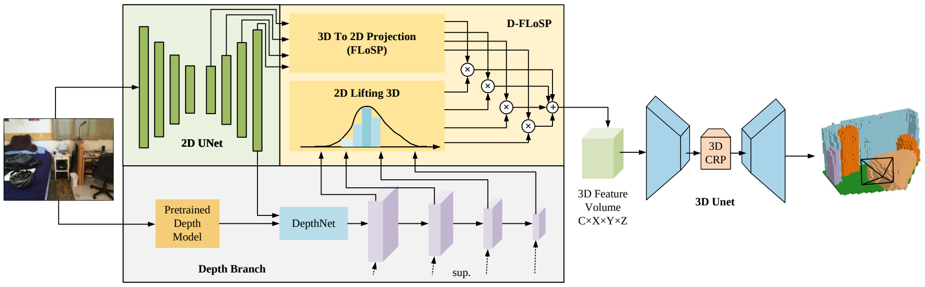


Fig. 1: The core design of **ISO** centers around the transformation of features from 2D to 3D spaces, encompassing the *Depth Branch* and the *D-FLoSP* module. A depth branch is initially integrated, it leverages a pre-trained depth model to estimate a pixel-wise depth map which is processed by the DepthNet to generate the final depth distribution. An element-wise multiplication between the voxel depth and features followed by summation are subsequently performed to derive the initial 3D voxel feature. The 3D feature is further processed to predict the 3D scene occupancy.

instance segmentation tasks. It predicts complete geometric reconstruction, semantic, and instance segmentations from a single RGB image's camera frustum. SCFusion [42], a real-time scene reconstruction framework, integrates continuous depth data using neural architecture for occupancy maps. It efficiently combines semantic completion with voxel states for simultaneous scene reconstruction and semantic understanding in real-time 3D environments.

3 Method

Problem definition. We focus on the problem of monocular 3D Occupancy Prediction. Specifically, this task takes a single RGB image $\mathbf{I}^{RGB}$ as input and output a voxel-wise occupancy along with semantic categories $\mathbf{Y}^{X \times Y \times Z \times C}$. X, Y, and Z represent the dimensions of the predicted 3D scene, while C denotes the total number of semantic categories.

Model overview. In this section, we introduce the overall architecture of ISO, as shown in Fig. 1. Following Monoscene [5], we utilize 2D Unet and 3D Unet architectures to handle 2D and 3D features. Our core design focuses on transforming features from 2D to 3D, incorporating the *Depth Branch* and *D-FLoSP* module. Specifically, we incorporate a depth branch to estimate the pixel-wise depth map using a pre-trained depth model. Processed by the DepthNet, the model outputs a refined depth distribution. Then, voxel depth and features are multiplied element-wise and summed to obtain the initial 3D voxel feature volume $\mathbf{X^{3d}} \in \mathbb{R}^{X \times Y \times Z \times C}$. Finally, after processing the 3D voxel features, the model outputs the 3D scene occupancy.

3.1 Depth Branch

In this section, we introduce how to effectively estimate depth information from a single image.

Coarse depth estimation. Learning depth from scratch can be quite challenging. However, thanks to the recent rapid advancements in depth estimation [3, 4, 27, 44], we can leverage pre-trained depth models to initially estimate a coarse depth map.

Compared to past models that could only estimate relative depth, we opted for Depth-Anything [44] as the pre-trained depth model, because it excels in predicting metric depth, as the Eq. (1) shows:

$$\mathbf{D}^{\text{metric}} = \mathbf{N}_{\text{depth}}(\mathbf{I}^{\text{rgb}}) \in \mathbb{R}^{1\times H\times W}, \tag{1}$$

where $\mathbf{N}_{\text{depth}}$ denotes the pre-trained depth model [3, 44], and $\mathbf{I}^{\text{rgb}}$ represents the input image. H and W represents the height and width of the input image.

Depth refinement. The coarse depth estimation is subsequently refined through model learning. The depth from pre-trained depth model is not precise enough for a higher mIoU, so we design a fine-tuning strategy. Specifically, the predicted metric depth $\mathbf{D}^{\text{metric}}$ is then concatenated with the image feature $\mathbf{X}$ that has the same spatial scale. The augmented feature is continue processed by a following DepthNet to get a refined depth distribution $\mathbf{D}^{\text{dist}}_{\text{s=1}}$, as Eq. (2) shows:

$$\mathbf{D}^{\text{dist}}_{\text{s=1}} = \mathbf{F}_{\text{depth}}\Big(\text{Concat}\big(\mathbf{D}^{\text{metric}},\ \mathbf{X}^{\text{2d}}_{\text{s=1}}\big)\Big),\ \mathbf{D}^{\text{dist}}_{\text{s=1}} \in \mathbb{R}^{N_{\text{bins}}\times H\times W}, \tag{2}$$

where $\mathbf{F}_{\text{depth}}$ is the DepthNet, $\mathbf{X}^{\text{2d}}_{\text{s=1}}$ is the 2D image feature, and N_{bins} is the number of discrete depth bins. s denotes the level of feature scale, where $s = 1$ implies that the feature is at a 1:1 original image ratio.

We down-samples the scale $\frac{1}{1}$ distribution to get other smaller-scale 2D image feature maps' depth distribution. We represent the procedure as Eq. (3):

$$\mathbf{D}^{\text{dist}}_{\text{s}=k} = \text{DownSample}(\mathbf{D}^{\text{dist}}_{\text{s=1}}),\ k = 2,\ 4,\ 8. \tag{3}$$

The refined depth is supervised by the ground truth $\mathbf{D}^{\text{GT}}$. The ground truth has single depth value at each pixel, so we convert the single value to a one-hot vector of length N_{bins}. The refined depth prediction can be optimized using the BCE loss function as Eq. (4) shows.

$$\begin{aligned}\mathcal{L}_{\text{depth}} = -&\frac{1}{N_{\text{bins}}\times H\times W}\\ &\sum_{d=1}^{N_{\text{bins}}}\sum_{h=1}^{H}\sum_{w=1}^{W}[\mathbf{D}^{\text{GT}}_{N_{\text{bins}},d,h,w}\cdot\log(\mathbf{D}^{\text{dist}}_{N_{\text{bins}},d,h,w})\\ &\qquad+(1-\mathbf{D}^{\text{GT}}_{N_{\text{bins}},d,h,w})\cdot\log(1-\mathbf{D}^{\text{dist}}_{N_{\text{bins}},d,h,w})]\end{aligned} \tag{4}$$

3.2 Dual Feature Line of Sight Projection (D-FLoSP)

Compared to the FLoSP module in MonoScene [5], which projects 3D voxels to 2D, and 3D feature vectors are the corresponding pixel feature vectors. It only considers the projection of 3D rays and overlooks depth information, our proposed D-FLoSP module can more effectively integrate depth information.

Specifically, after getting depth distribution in the camera frame, we implement the FLoSP module to depth distribution. Each 3D voxel centroid position x^c in the world frame can be projected to the camera frame and the pixel frame using camera pose and intrinsic matrix. We assign the projected depth values to discrete depth bins index by Eqs. (5) and (6), according to [32],

$$\delta = \frac{2(d_{\max} - d_{\min})}{N_{bins}(1 + N_{bins})}, \tag{5}$$

$$l = -0.5 + 0.5\sqrt{1 + \frac{8(d - d_{\min})}{\delta}}, \tag{6}$$

and the depth distribution probability it lies in serves as the depth probability of that voxel. The process is illustrated by Eq. (7):

$$\mathbf{D}^{3\mathrm{d}}_{\mathrm{s}=k} = \mathbf{\Phi}^{3\mathrm{d}}_{\rho(x^c)}(\mathbf{D}^{\mathrm{dist}}_{\mathrm{s}=k}) \in \mathbb{R}^{X\times Y\times Z\times C}, k = 1, 2, 4, 8, \tag{7}$$

where $\mathbf{\Phi}^{3\mathrm{d}}_a(b)$ is the 3D sampling of b at coordinates a, and $\rho(\cdot)$ is the perspective projection. Thus, the depth distribution is lifted from 2D to 3D. The bin-based representation with quantization is used for mapping continuous depth value to a discrete form, each bin will be assigned a different depth distribution score to make voxel aware of depth. The original depth precision is a continuous value, and hard to integrate it with the discrete form of voxel.

Following [5], we also employ the FLoSP module to generate a 3D voxel feature from a 2D feature map. Similar to what we did with the voxel centroids, the projected voxel can sample corresponding 1:k scaled feature maps $F_{1:k}$ from the 2D UNet decoder. This process is illustrated by Equation Eq. (8):

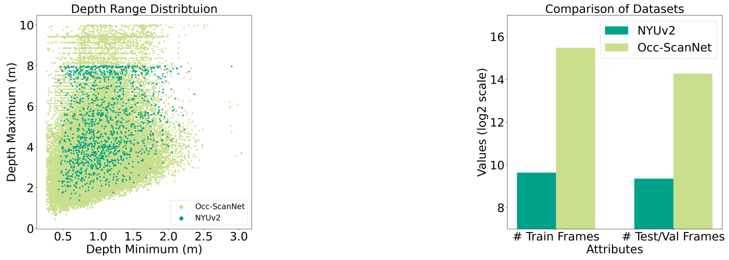
$$\mathbf{X}^{3\mathrm{d}}_{\mathrm{s}=k} = \mathbf{\Phi}^{2\mathrm{d}}_{\rho(x^c)}(\mathbf{X}^{2\mathrm{d}}_{\mathrm{s}=k}) \in \mathbb{R}^{X\times Y\times Z\times C}, k = 1, 2, 4, 8, \tag{8}$$

where $\mathbf{\Phi}^{2\mathrm{d}}_a(b)$ is the 2D sampling of b at coordinates a, and $\rho(\cdot)$ is the perspective projection.

The final 3D feature map $\mathbf{X}^{3\mathrm{d}}$ are summed in Eq. (9):

$$\mathbf{X}^{3\mathrm{d}} = \sum_{\mathrm{s}\in\{1,2,4,8\}} \mathbf{X}^{3\mathrm{d}}_{\mathrm{s}=k} \odot \mathbf{D}^{3\mathrm{d}}_{\mathrm{s}=k}, \tag{9}$$

where $\odot$ is the element-wise multiplication. The output map $\mathbf{X}^{3\mathrm{d}}$ is then serves as the input of 3D UNet.



(a) Depth ranges in NYUv2 and our Occ-ScanNet Benchmark

(b) Dataset attributes of NYUv2 and our Occ-ScanNet Benchmark

Fig. 2: Comparison of NYUv2 and Occ-ScanNet Benchmark. In (a), the depth ranges of NYUv2 and Occ-ScanNet are distinguished by dark and light green, respectively, with the horizontal axis indicating the minimum depth and the vertical axis showing the maximum depth of scenes. (b) quantitatively demonstrates that Occ-ScanNet possesses a significantly larger data scale compared to the NYUv2 dataset.

4 Occ-ScanNet Benchmark

Compared to the previous widely used indoor scene benchmark NYUv2 [29], which contains only 795 / 654 for train / test samples, our benchmark boasts 45,755 / 19,764 samples. Our benchmark significantly surpasses NYUv2 [29] in both data quantity and richness of scene depth, as illustrated in Fig. 2. This dataset is available at https://huggingface.co/datasets/hongxiaoy/OccScanNet.

4.1 Overview

Occ-ScanNet benchmark features a train/validation split of 45,755 / 19,764 samples. As shown in Fig. 3, Occ-ScanNet exhibits rich diversity in scenes and viewpoints. This diverse dataset not only challenges the task of predicting scene occupancy but also fosters future research endeavors towards developing larger-scale and more versatile occupancy models.

4.2 Occupancy Label Generation

We adhere to the data formulation used in NYUv2 [29] dataset. The pipeline of occupancy generation process is showed in Fig. 4. For generating 3D voxel labels, we initially followed the generation process in CompleteScanNet and then employed manual inspection.

More specifically, from the official ScanNet [8] scenes, we extracted various data components, including color images, depth images, camera intrinsic matrices, and camera poses. A total of 100 frames were sampled from each scene and were randomly divided into training and validation sets with a 7/3 split ratio per scene.

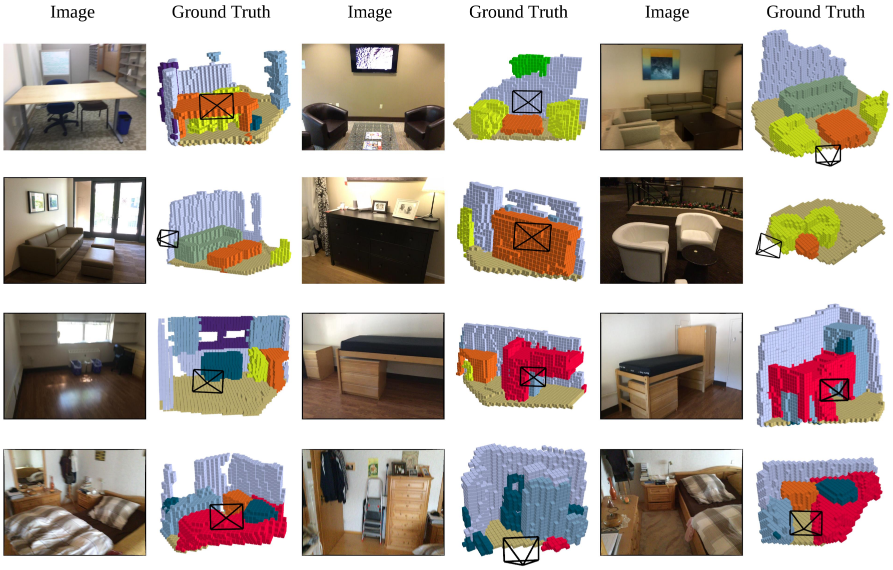


Fig. 3: Samples Visualization in Occ-ScanNet Benchmark. The original RGB image is shown in column 1,3 and 5, the corresponding scene voxel labels is shown in column 2, 4 and 6. The first two rows are different views from different scenes and the last two rows each is three different views from the same scene.

Manual inspection. The manual inspection involves verifying camera poses, camera intrinsics, and voxel positions. We filter out erroneous samples to prevent convergence issues in the model training process.

Voxel label assignment. For each frame, only a specific area in front of the camera is defined for analysis. Therefore, the selection of the voxel origin is paramount, as it determines the subsequent coordinates of other voxel. Subsequently, each voxel is assigned a label based on its nearest voxel in the CompleteScanNet dataset. Additionally, we excluded frames with a ratio of unknown or empty labels exceeding 95% or frames where the number of semantic label classes was less than two. Consequently, we obtained the generated 3D voxel labels for each frame.

5 Experiments

5.1 Setup

Datasets. The NYUv2 [29] dataset provides scenes represented in 240 × 144 × 240 voxels grids, labeled with 13 classes, including 1 for free space, 1 for unknown,

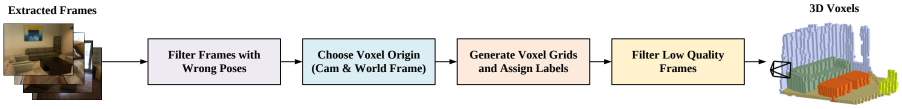

Extracted Frames
Filter Frames with Wrong Poses
Choose Voxel Origin (Cam & World Frame)
Generate Voxel Grids and Assign Labels
Filter Low Quality Frames
3D Voxels

Fig. 4: Pipeline of Occ-ScanNet dataset label generation. Color images, depth images, camera intrinsic and poses are extracted from ScanNet scenes. For each scene, 100 frames were sampled and randomly split into training and validation sets with a 7/3 ratio. Frames with invalid camera poses or exceeding scene boundaries were filtered out. Only the area in front of the camera was analyzed, necessitating careful selection of the voxel origin. Voxel were labeled based on their nearest voxel in the CompleteScanNet dataset. Frames with >95% unknown/empty labels or <2 semantic classes were excluded, resulting in generated 3D voxel labels for each frame.

and 11 for specific semantics (ceiling, floor, wall, window, chair, bed, sofa, table, tvs, furniture, objects). The dataset consists of 795 / 654 scenes in the train / test splits. The model is trained and evaluated on down-sampled $60 \times 36 \times 60$ voxels.

Occ-ScanNet dataset provides scenes represented in $60 \times 60 \times 36$ voxel grids, labeled with 12 classes including 1 for free space, and 11 for specific semantics (ceiling, floor, wall, window, chair, bed, sofa, table, tvs, furniture, objects). The dataset comprises 45,755 / 19,764 frames in the train / val splits. The model is trained and evaluated on the original scale.

Occ-ScanNet-mini dataset has the same class number and voxel scene size as Occ-ScanNet, except that this mini dataset consists of 4,639 / 2,007 frames in the train / val splits. The model is also trained and evaluated on the original scale.

Implementation Details. We employ a pre-trained EfficientNet-B7 [31] as the encoder in our 2D UNet architecture. In the depth branch, we utilize a pre-trained DepthAnything model [44], which remains frozen during training. Additionally, we integrate a depth loss specific to the depth branch, complementing the other losses outlined in Monoscene [5]. Our model is trained on two datasets: NYUv2 [29] and Occ-ScanNet. For NYUv2, ISO undergoes 30 epochs of training using AdamW optimizers. Initially, the learning rate is set to 5e-6 for the DepthNet and 1e-4 for other components, with a learning rate reduction by 0.1 at epoch 20. Training on NYUv2 takes approximately 7 hours using 2 NVIDIA L20 GPUs (2 items per GPU). On the Occ-ScanNet-mini dataset, ISO is trained for 60 epochs under similar learning rate settings, decreasing the rate by 0.1 at epoch 40. This training process takes around 12 hours using 8 L20 GPUs. For Occ-ScanNet, ISO is trained for 10 epochs with the same learning rate schedule as used for Occ-ScanNet-mini, requiring approximately a day using 8 L20 GPUs.

5.2 Main Results

Table 1: Performance on the Occ-ScanNet dataset

Method	Input	IoU	ceiling 🟥	floor 🟩	wall 🟦	window 🟦	chair 🟨	bed 🟧	sofa 🟪	table 🟦	tvs 🟩	furniture 🟧	objects 🟪	mIoU
MonoScene* [5]	x^{rgb}	41.60	15.17	**44.71**	**22.41**	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84	24.62
ISO(Ours)	x^{rgb}	**42.16**	**19.88**	41.88	22.37	**16.98**	**29.09**	**42.43**	**42.00**	**29.60**	**10.62**	**36.36**	**24.61**	**28.71**

Occ-ScanNet performance. We first evaluate our model's performance on the large-scale Occ-ScanNet dataset. As shown in Tab. 1, the results indicate that our method significantly outperforms MonoScene [5]. The $*$ denotes results obtained using their code trained on our dataset. In Fig. 5, we also conducted qualitative visualization comparisons.

NYUv2 performance. The NYUv2 [29] dataset serves as a widely used benchmark for indoor scene evaluation. The results in Tab. 2 demonstrate that our method achieves state-of-the-art performance. Our method also demonstrates substantial performance improvement on the NYUv2 dataset, in addition to the gains observed on Occ-ScanNet. As illustrated in Fig. 6, our ISO demonstrates superior room layout prediction compared to [5], particularly evident in rows 1 and 2. This improvement can be attributed to incorporating depth knowledge into the model. Additionally, our method exhibits the capability of learning small objects, such as pictures on the wall in row 3.

Table 2: Performance on the NYUv2 dataset

Method	Input	IoU	ceiling 🟥	floor 🟩	wall 🟦	window 🟦	chair 🟨	bed 🟧	sofa 🟪	table 🟦	tvs 🟩	furniture 🟧	objects 🟪	mIoU
LMSCNet$^{\text{rgb}}$ [28]	$\hat{x}^{\text{occ}}$	33.93	4.49	88.41	4.63	0.25	3.94	32.03	15.44	6.57	0.02	14.51	4.39	15.88
AICNet$^{\text{rgb}}$ [18]	x^{rgb}, $\hat{x}^{\text{depth}}$	30.03	7.58	82.97	9.15	0.05	6.93	35.87	22.92	11.11	0.71	15.90	6.45	18.15
3DSketch$^{\text{rgb}}$ [6]	x^{rgb}, $\hat{x}^{\text{TSDF}}$	38.64	8.53	90.45	9.94	5.67	10.64	42.29	29.21	13.88	9.38	23.83	8.19	22.91
MonoScene [5]	x^{rgb}	42.51	8.89	93.50	12.06	12.57	13.72	48.19	36.11	15.13	15.22	27.96	12.94	26.94
NDC-Scene [45]	x^{rgb}	44.17	12.02	**93.51**	13.11	13.77	15.83	49.57	39.87	17.17	24.57	31.00	14.96	29.03
ISO(Ours)	x^{rgb}	**47.11**	**14.21**	93.47	**15.89**	**15.14**	**18.35**	**50.01**	**40.82**	**18.25**	**25.90**	**34.08**	**17.67**	**31.25**

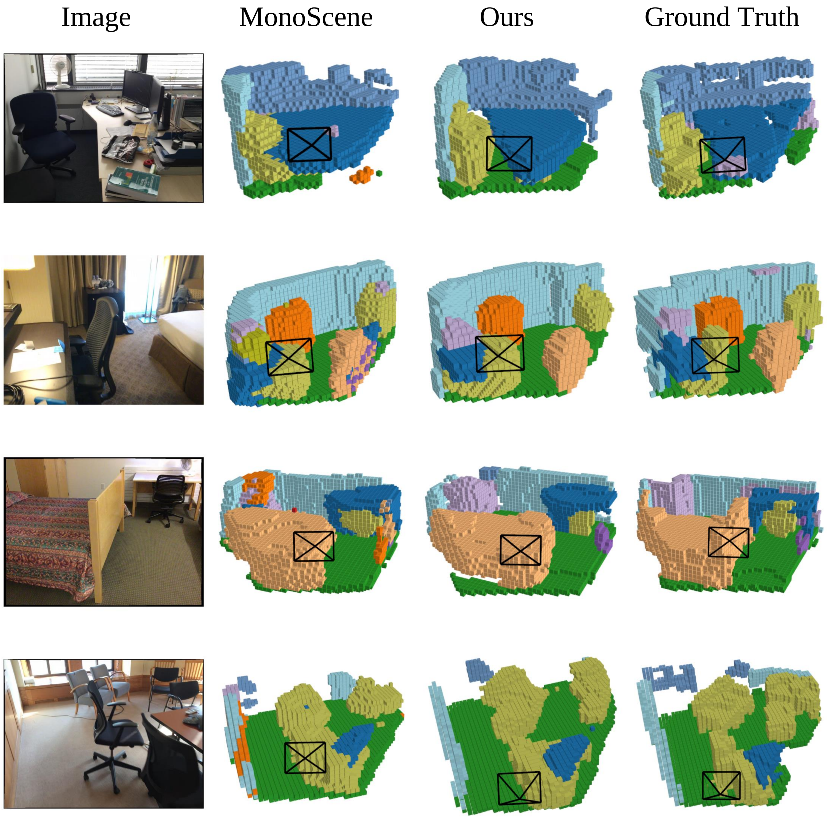


Fig. 5: Qualitative Analysis on the Occ-ScanNet Dataset. The input image is displayed on the left, while the predicted scene is shown in the middle two column, and ground truth on the right column.

5.3 Ablation Study

We conduct ablation studies on the NYUv2 [29] and Occ-ScanNet-mini datasets to validate the effectiveness of our model design.

Depth fusion. Inspired by BEVDepth [20], we initially adopt the integration of BEV depth information, denoted as '+bev-depth'. However, we observed that while BEV depth works well for 3D detection in outdoor scenes, directly applying it to indoor scenes is less effective. We speculate that this discrepancy arises because height information is more crucial in indoor scenes, where indoor heights often correspond to different 3D structures and semantics. Therefore, we adopt voxel depth fusion, denoted as '+voxel-depth'. Specifically, we project each 3D voxel centroid to a frustum grid to sample the depth distribution probability

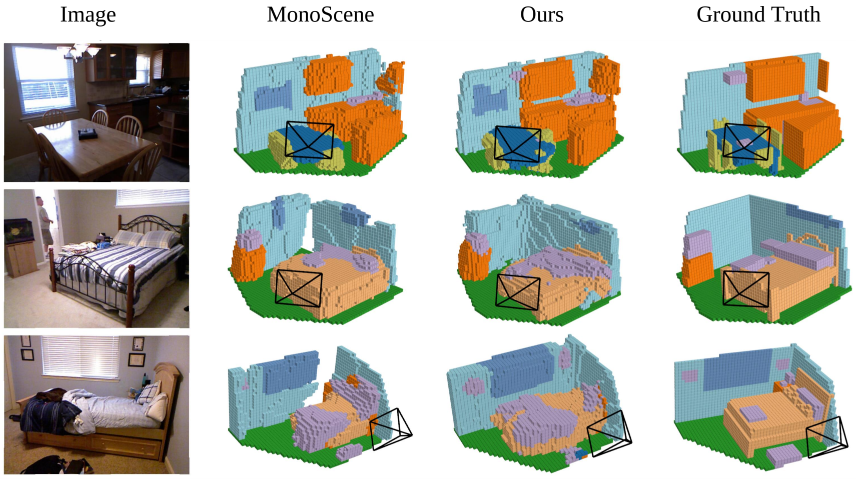


Fig. 6: Qualitative Analysis on the NYUv2 Dataset: The input is displayed on the left, with the camera viewing frustum shown in every image. ISO adeptly captures the scene layout and delineates various semantic instances more accurately. Notably, it excels in reconstructing the corners of walls (row 2) and accurately representing small objects like pictures on the wall (row 3).

of the voxel. We then weight the 2D feature at that pixel coordinate as the voxel feature. The results in Tab. 3 further validate that the voxel depth fusion approach is more effective.

The BEV method usually performs better in detecting ceilings and floors but performs poorly in detecting furniture such as windows, walls, televisions, beds, sofas, and tables. Ceilings and floors usually have relatively flat and regular geometric shapes, which makes their projections clearer and more consistent from a BEV perspective. In contrast, objects such as windows, walls, and televisions typically have more complex geometric shapes. In the BEV perspective, the projections of these objects may become blurry or overlapping, making it difficult to accurately distinguish and detect.

Table 3: Ablation study of depth fusion on the NYUv2 and Occ-ScanNet-mini dataset

Method	Input	NYUv2		Occ-ScanNet-mini	
		IoU	mIoU	IoU	mIoU
baseline	x^{rgb}	42.27	27.13	50.94	38.95
+ bev-depth	x^{rgb}	42.67	27.14	**51.58**	38.48
+ voxel-depth	x^{rgb}	**47.11**	**31.25**	51.03	**39.08**

Table 4: Ablation of depth information on the NYUv2 dataset

depth-method	learned	multi-scale	IoU	mIoU
GT		✓	53.98	34.47
ZoeDepth	✓		45.18	29.15
ZoeDepth	✓	✓	45.24	29.40
DepthAnything			44.77	29.28
DepthAnything		✓	45.48	29.57
DepthAnything	✓	✓	46.94	31.02
DepthAnything	✓		**47.11**	**31.25**

Different depth pre-trained models. We conducted ablation experiments on the depth method using the Occ-ScanNet-mini dataset. Initially, we observed the upper bound of model performance when ground truth (GT) depth information was provided. Subsequently, we initialized our depth module using ZoeDepth [3] and Depth-Anything [44]. Experimental results indicate that utilizing Depth-Anything [44] yields better performance. Moreover, our approach allows for further enhancement by fine-tuning a pre-trained depth module, resulting in improved performance.

Multi-scale depth. In Tab. 4, we investigate the impact of using multi-scale depth distribution on the NYUv2 [29] dataset. Specifically, the multi-scale depth distribution is used to weight the multi-scale 3D voxel feature. Meanwhile, we tried a method that only uses the single-scale depth distribution map to generate the 3D voxel depth matrix. The used depth distribution map has a scale of $\frac{1}{8}$ original image scale. Only this single-scale 3D voxel matrix is used to weigh the summed-up 3D voxel feature out of the FLoSP module. Without multi-scale depth, all 3D features are simply added up and a single depth distribution is used for weighting. This method ignores the different importance that features at different scales may have, as all features at all scales are treated equally and all depth information is weighted by a unified depth distribution. With multi-scale depth, the 3D features of each scale are first weighted with the depth distribution of their corresponding scale, and then these weighted features are added up. This method takes into account the differences in the importance of features at different scales, as each scale's feature is weighted based on its corresponding depth distribution.

In real-world scenarios, various objects and structures (such as ceilings, walls, furniture, etc.) often exhibit different scales. Hence, accounting for scale variations is crucial for accurately modeling these objects. Multi-scale depth can effectively capture these scale changes by assigning different weights to features at each scale. Depth information is indispensable for 3D scene completion. Leveraging the depth distribution information at each scale allows us to weight the corresponding 3D features, enabling the model to better understand and represent the objects in the scene and their relationships.

5.4 Discussion

Data scaling up. In Tab. 5, we delve into the impact of data volume on our method's performance by scaling up the number of scene samples. We trained the model on Occ-ScanNet using 10% and 100% scene samples respectively, and tested the model performance. The results indicate that larger dataset exhibits more significant gains.

Table 5: Scaling up comparison on the Occ-ScanNet

Data	Depth Branch	Ours	
		IoU	mIoU
10%	✓	21.79	9.57
100%	✓	42.16	28.71

Limitations and future works. Although our model can estimate 3D occupancy effectively with the assistance of depth information, semantic learning still faces challenges due to class imbalance issues. Furthermore, our proposed Occ-ScanNet only considers 11 common semantic classes, which may not fully capture the diversity of categories present in real-world scenarios. Despite the significant increase in data volume compared to previous datasets, Occ-ScanNet remains limited in the number of scenes it covers. In future work, we will focus more on semantic exploration and conduct tests in more generalized scenarios.

6 Conclusion

In this paper, we introduce ISO, a novel method for monocular 3D occupancy prediction. ISO utilizes a D-FLoSP (Dual Feature Line of Sight Projection) and a multi-scale feature fusion strategy to address variations in scene and object sizes, specifically tailored for indoor scenes. Additionally, we present a new benchmark, Occ-ScanNet, aimed at fostering scalable studies of indoor scenes. We hope that this dataset will attract more attention to research on indoor occupancy prediction. While achieving promising results in 3D structure prediction, accurately identifying semantics remains a significant challenge in the future.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0160102), the National Natural Science Foundation of China (No. U21B2042, No. 62320106010), and in part by the 2035 Innovation Program of CAS, and the InnoHK program.

References

1. Arshad, M.S., Beksi, W.J.: List: Learning implicitly from spatial transformers for single-view 3d reconstruction. In: ICCV (2023)
2. Azuma, R.T.: A survey of augmented reality. Presence: teleoperators & virtual environments (1997)
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
4. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
5. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: CVPR (2022)
6. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: CVPR (2020)
7. Dahnert, M., Hou, J., Nießner, M., Dai, A.: Panoptic 3d scene reconstruction from a single rgb image. NeurIPS (2021)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
9. Denninger, M., Triebel, R.: 3d scene reconstruction from a single viewport. In: ECCV (2020)
10. DeSouza, G.N., Kak, A.C.: Vision for mobile robot navigation: A survey. TPAMI (2002)
11. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: CVPR (2017)
12. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Reconstructing three-dimensional models of interacting humans. arXiv preprint arXiv:2308.01854 (2023)
13. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023)
14. Guan, H., Song, C., Zhang, Z.: Gramo: geometric resampling augmentation for monocular 3d object detection. Frontiers of Computer Science (2024)
15. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
16. Huang, Y., Zheng, W., Zhang, B., Zhou, J., Lu, J.: Selfocc: Self-supervised vision-based 3d occupancy prediction. In: CVPR (2024)
17. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: CVPR (2023)
18. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: CVPR (2020)
19. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: CVPR (2023)
20. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: AAAI (2023)
21. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022)

22. Li, Z., Yu, Z., Austin, D., Fang, M., Lan, S., Kautz, J., Alvarez, J.M.: Fb-occ: 3d occupancy prediction based on forward-backward view transformation. arXiv preprint arXiv:2307.01492 (2023)
23. Liu, S., Hu, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., Li, X.: See and think: Disentangling semantic scene completion. NeurIPS (2018)
24. Murez, Z., Van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images. In: ECCV (2020)
25. Pan, M., Liu, J., Zhang, R., Huang, P., Li, X., Liu, L., Zhang, S.: Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. arXiv preprint arXiv:2309.09502 (2023)
26. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR (2019)
27. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. TPAMI (2020)
28. Roldao, L., de Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 3DV (2020)
29. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
30. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
31. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
32. Tang, Y., Dorn, S., Savani, C.: Center3d: Center-based monocular 3d object detection with joint depth understanding. In: DAGM GCPR (2020)
33. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. NeurIPS (2023)
34. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: ICCV (2023)
35. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: ICCV (2023)
36. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: ICCV (2023)
37. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: CoRL (2022)
38. Wang, Y., Chen, Y., Liao, X., Fan, L., Zhang, Z.: Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In: CVPR (2024)
39. Wang, Y., Chen, Y., Zhang, Z.: Frustumformer: Adaptive instance-aware resampling for multi-view 3d detection. In: CVPR (2023)
40. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV (2023)
41. Wu, Q., Wang, K., Li, K., Zheng, J., Cai, J.: Objectsdf++: Improved object-compositional neural implicit surfaces. In: ICCV (2023)
42. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Scfusion: Real-time incremental scene reconstruction with semantic completion. In: 3DV (2020)
43. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., et al.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In: CVPR (2023)

44. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
45. Yao, J., Li, C., Sun, K., Cai, Y., Li, H., Ouyang, W., Li, H.: Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In: ICCV (2023)
46. Yu, Z., Shu, C., Deng, J., Lu, K., Liu, Z., Yu, J., Yang, D., Li, H., Chen, Y.: Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. arXiv preprint arXiv:2311.12058 (2023)
47. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE access (2020)
48. Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: ECCV (2018)
49. Zhang, X., Bi, S., Sunkavalli, K., Su, H., Xu, Z.: Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In: CVPR (2022)
50. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: ICCV (2019)
51. Zhong, M., Zeng, G.: Semantic point completion network for 3d semantic scene completion. In: ECAI (2020)