2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Multi-scale Depth-Calibrated Kernel-split Network for Monocular Occupancy Prediction

Xiaoke Tan[1], Jinlai Zhang[2](✉), Yuxin Chen[3], Kai Gao[2], Qiqi Li[2], Lin Hu[2] and Gang Wu[1,2]

[1] Elite Engineering School, Changsha University of Science and Technology,
Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
jinlai.zhang@csust.edu.cn
[3] School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha, 410114, Hunan, China

Abstract. In the task of indoor occupancy prediction from a single image, there is an issue where it is challenging to complement the semantic scene due to the small size and intricate nature of the predicted objects. Traditional methods struggle to obtain rich and accurate semantic information, which results in the worsening of the problem of semantic scene blurring. To solve the problem, we propose a novel approach accompanied by our proposed Multi-scale Context Calibration Module (MCCM), Depth Calibrated Residual Network (DCRNet) and Kernel-split Depthwise Attention (KSDA) to enhance the scene semantic information and alleviate the depth blurring problem of semantic scenes. Ablation experiments confirm the effectiveness of our module. Comparative analysis with the SOTA model verifies the superiority and generalisation of our model. Comparison on the OccScanNet_mini dataset confirms the excellent generalisation of our method even with limited data. Specifically, our method reaches 47.94% and 32.33% for IoU and mIoU on the NYUv2 dataset, and 42.81% and 29.59% for IoU and mIoU on the OccScanNet dataset, respectively.

Keywords: Indoor occupancy prediction, Semantic scene completion, Computer vision.

1 Introduction

The task of 3D semantic scene complementation (SSC) from a single RGB image plays a pivotal role in the domain of computer vision and 3D scene understanding. Its objective is to deduce the complete geometric structure of a 3D semantic scene from a single

RGB image and to assign accurate semantic labels to each 3D voxel. In recent years, the rapid advancements in deep learning techniques and the availability of large-scale 3D datasets have significantly propelled the development of 3D semantic segmentation[6, 18, 13, 8, 23, 11] and depth estimation[9, 15, 20, 14]. Nevertheless, despite the substantial progress in outdoor semantic scene complementation, there remains a pressing demand for an efficient method capable of performing indoor semantic scene complementation. Such a method would be essential for inferring a complete and semantically rich 3D scene from a single RGB image.

In contrast to methods that rely on depth sensors[12, 7, 5, 19, 1], 3D semantic scene complementation using a single RGB image only requires a standard camera, which is hardware-lightweight and more conducive to consumer-grade applications. However, the absence of stereo visual cues in such models results in scale ambiguity in depth prediction and errors in geometric reconstruction, particularly in complex occlusion scenarios. The existing method, MonoScene[2], achieves end-to-end complementation from monocular RGB images to dense 3D semantic scenes for the first time, overcoming the reliance of traditional methods on depth input. NDCScene[21] extends this approach by introducing the normalized device coordinate (NDC) space, which helps address feature size ambiguity and depth-related feature ambiguity. Additionally, the Depth Adaptive Dual Decoder (DADD) is introduced to tackle challenges associated with the use of a single decoder in the target 3D space, which makes multi-scale 2D and 3D feature fusion difficult and fails to adequately consider depth perception. ISO[22] significantly improves the handling of indoor scenes by employing the Depth-Anything model to generate metric depth maps, performing multi-scale[28] feature fusion, and introducing a novel projected dual-feature-implemented projection (D-FloSP) mechanism. This approach is designed to effectively manage the complexity and object size diversity inherent in indoor scenes. While the aforementioned ISO methods have demonstrated strong performance in the task of 3D semantic scene complementation from a single RGB image, establishing them as the state-of-the-art (SOTA) model for this task, they suffer from limitations. Specifically, the lack of dynamic calibration in the output of the 2D encoder results in poor multi-scale feature fusion, inadequate processing of depth estimation information, an inability to obtain precise depth distributions, and insufficient depth inference in occluded regions, leading to depth blurring.

In this paper, we present a novel framework for monocular occupancy prediction that integrates multi-scale semantic context calibration with depth information. Our approach is designed to address the key challenges of small object detection, semantic scene blurring, and depth misalignment, which are common in traditional monocular methods. To achieve this, we propose a Multi-scale Context Calibration Module (MCCM), Depth Calibrated Residual Network (DCRNet), and Kernel-split Depthwise Attention (KSDA). These components work synergistically to enrich the semantic understanding of indoor scenes and alleviate the depth blurring problem inherent in monocular vision. Among them, MCCM enhances the multi-scale key feature information for the 2D encoder. DCRNet is used to fully optimise the depth distribution information of the semantic scene. Finally, KSDA is used to efficiently capture long range dependencies while reducing the information loss caused by downsampling operations in the 3D encoder. The core contributions of this paper are summarized as follows:

- We propose the MCCM for calibrating multi-scale features and retaining key features.
- We propose the DCRNet for refining depth information and alleviating the depth ambiguity problem.
- We propose the KSDA module for efficiently capturing long-distance dependencies and enhancing depth information while recovering the fine geometry of the scene.

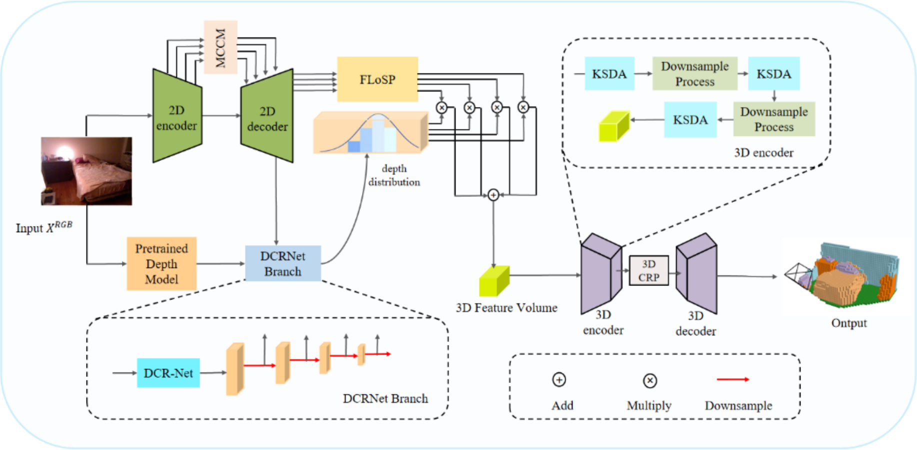

Fig. 1. The overview of the proposed MDKOcc.

2 Methodology

In the indoor occupancy prediction tasks, significant challenges arise from scene semantic information scarcity and depth ambiguity. Compared to outdoor scenarios, this task is further characterized by the need to predict objects with smaller dimensions and greater categorical diversity. These challenges collectively amplify the difficulty of achieving accurate semantic scene completion. This paper proposes an innovative methodology termed **M**ulti-scale **D**epth-Calibrated **K**ernel-split Network for Monocular **Occ**upancy Prediction (MDKOcc) to address these issues. Specifically, we introduce three core components, including Multi-scale Context Calibration Module (MCCM), Depth Calibrated Residual Network (DCRNet), and Kernel-split Depthwise Attention (KSDA). These modules synergistically enhance scene semantics, mitigate depth estimation uncertainties, while simultaneously strengthening multi-scale feature representation in 2D encoders and preserving fine-grained detail information in 3D shallow encoders. The comprehensive architectural schematic is illustrated in Fig.1. The following section describes the proposed modules one by one.

2.1 Multi-scale Context Calibration Module

In the process of performing 2D visual feature extraction, 2D Unet passes encoder features directly to the decoder through jump connections, but the low-level features[24-27] contain noise and redundant information. This not only leads to noise amplification of the model as well as flooding of the extracted multi-size key features, resulting in the lack of key feature information, which affects the processing effect of the subsequent 3D feature projection as well as the refinement of the depth information, but also makes the 3D decoder lack of sufficient semantic information, which in turn exacerbates the problem of depth ambiguity.

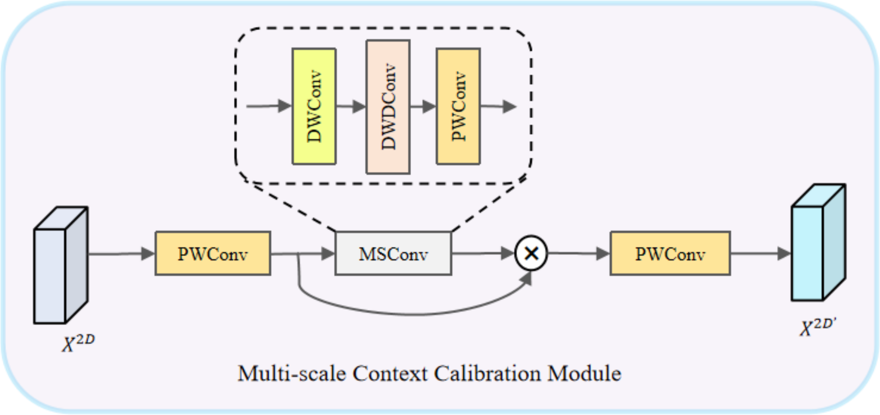

Fig. 2. Detailed structure of MCCM.

In order to solve the above problem, we designed Multi-scale Context Calibration Module for calibrating the 2D encoder's output to retain important features and suppress irrelevant information. The specific structure of the module is shown in Fig.2. We can see that the output features are firstly adjusted by a PWConv. The feature processing is then performed via Multi-scale Conv (MSConv), which specifically includes the use of a 7×7 DWConv to capture local to medium-range contextual information, and a 9×9 DWDConv to extract long-range dependencies and strengthen the decoder's ability to understand the global structure. The channel weights are dynamically adjusted using PWConv to enhance key features. Finally, the features before and after MSConv processing are fused and output after adjustment by PWConv. With this approach, we retain the key feature information in the encoder, enhance the subsequent 3D feature projection as well as depth information refinement, and alleviate the problem of depth ambiguity. At the same time, the model's understanding of the global context will be enhanced by extracting the long-range dependencies, making the model more accurate in complementing the occluded regions and improving the performance of the network. The structure of the module can be represented as:

$$X^{2D\prime} = F_{pw}(F_{pw}(X^{2D}) * F_{ms}(F_{pw}(X^{2D}))) \tag{1}$$

where X^{2D} is the input tensor and $X^{2D'}$ is the output tensor. F_{pw}, F_{ms} represent the operations PWConv, and MSConv respectively.

2.2 Depth Calibrated Residual Network

In the indoor occupancy prediction task, the ambiguity of depth information is a great challenge in this task. The original DepthNet optimises the coarse depth information generated by a pre-trained depth model and fuses it with 2D image features (X^{2D}) to generate a more accurate depth distribution. However, DepthNet is weak in refining the depth information sufficiently, resulting in inaccurate depth distributions obtained for calculating the depth probabilities after projecting the 3D voxel centres to the pixel coordinate system.

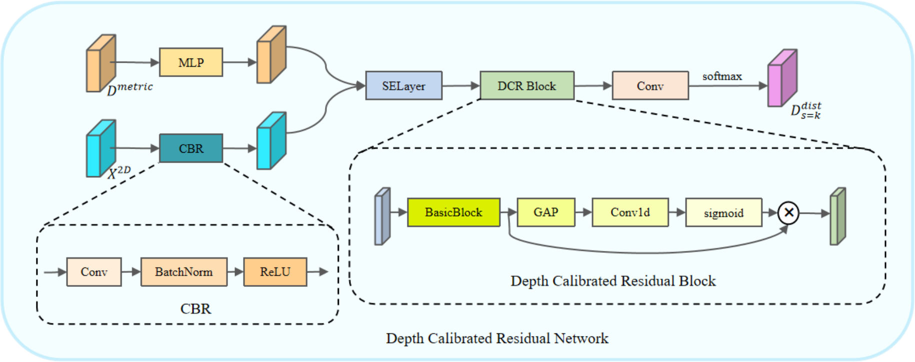

Fig. 3. Detailed structure of DCRNet.

In order to solve the above problem, we designed Depth Calibrated Residual Network (DCRNet), which is used to fully optimise the depth information and improve the accuracy of calculating the depth probability after projecting the centre of 3D voxels to the pixel coordinate system. The structure of the module is shown in Fig.3. It can be seen that firstly, the MLP operation is performed on the D^{metric}, and then secondly with the image features that have been processed by the CBR module processed X^{2D} after SELayer fusion, and then after Depth Calibrated Residual Block (DCR Block), the recalibration of the features is performed after residual processing. Finally, the output channel is adjusted by convolution to obtain the depth distribution after efficient refinement. This method performs the generation of weights as well as feature recalibration on the residual processed features as compared to the original method. In this way, DCRNet processes the depth information efficiently, alleviates the problem of ambiguous depth information to some extent, and improves the overall performance of the network. The module can be represented as:

$$D_{s=k}^{dist} = softmax(F_{conv}(F_{dcr}(F_{se}(F_{mlp}(D^{metric}), F_{cbr}(X^{2D}))))) \tag{2}$$

where F_{mlp}, F_{mlp}, F_{se}, F_{dcr}, F_{conv}, softmax represent the operations MLP, CBR, SE-Layer, DCR Block, Conv2d, and softmax respectively. D^{metric}, X^{2D} represent the inputs to DCRNet, and $D_{s=k}^{dist}$ represents the depth distribution after the refinement process.

2.3 Kernel-split Depthwise Attention

As 3D data is required to model the spatial continuity between voxels, the local convolution of the shallow 3D encoder is difficult to capture long distance dependencies, and the lack of depth inference for occluded regions makes the voxel prediction confidence in deeply ambiguous regions decrease. Meanwhile, in indoor scenes, the details of objects may be over-smoothed in shallow downsampling, resulting in the downsampling of the shallow encoder often losing the detail information, making it difficult for the decoder to recover the fine geometry. In addition, in the indoor occupancy prediction task, the scale complexity of the scene with large differences in object sizes makes it difficult for the model to accurately complement the semantic scene.

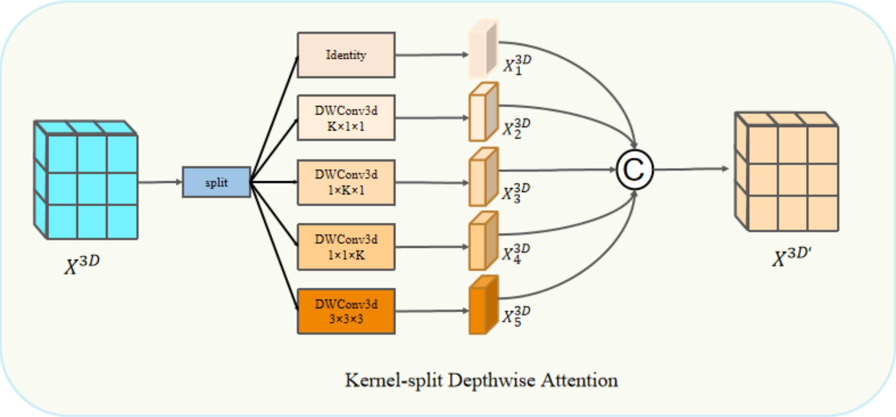

Fig. 4. Detailed structure of KSDA.

To solve the above problems, we designed Kernel-Split Depthwise Attention for efficiently capturing long range dependencies. At the same time, it reduces the missing information during the downsampling operation, and the structure of this module is shown in Fig.4. It can be seen that, due to the complexity of the large kernel convolution calculation in the 3D scene, we slice the input feature voxels, and then process them by using four DWConv, and finally, we concat five feature voxels. This method not only captures long range dependencies efficiently, but also reduces the number of parameters and computational effort while maintaining a large receptive field, avoiding the problem of cubic-level computational complexity growth in 3D scenarios. In addition, four DWConv obtain feature voxels at different scales, which are finally fused to enhance the richness of feature semantics. By using the method in conjunction with downsampling, it not only solves the problem of the shallow encoder's difficulty in capturing long-range dependencies and avoids the complexity of cubic-level computation in 3D scenarios, but also improves the model's ability of extracting wide-area contextual and edge information, and enhances the depth information of the network. Overall, the method mitigates the problems caused by the complexity of the scene scale and the large differences in object sizes. While maintaining computational efficiency, it also solves the problem of shallow encoders losing detail information during downsampling. The principle of the module can be expressed by the following equation:

$$X_{hwd}, X_{wd}, X_{hd}, X_{hd}, X_{id} = split(X^{3D}) \tag{3}$$

$$X_1^{3D} = X_{id} \tag{4}$$

$$X_2^{3D} = DWConv^{c}_{K\times1\times1}(X_{hw}) \tag{5}$$

$$X_3^{3D} = DWConv^{c}_{1\times K\times1}(X_{hd}) \tag{6}$$

$$X_4^{3D} = DWConv^{c}_{1\times1\times K}(X_{wd}) \tag{7}$$

$$X_5^{3D} = DWConv^{c}_{3\times3\times3}(X_{hwd}) \tag{8}$$

$$X^{3D'} = Concat(X_1^{3D}, X_2^{3D}, X_3^{3D}, X_4^{3D}, X_5^{3D}) \tag{9}$$

where c represents the number of channels in DWConv and K denotes the kernel size of the DWConv.

3 Experiments

3.1 Experimental setup

Dataset. In this paper, we conducted experiments using the NYUv2[17] dataset and the OccScanNet[22] dataset to validate the effectiveness of our method for the indoor monocular occupancy prediction task. The NYUv2 dataset contains a total of 1,449 RGB images containing 464 different indoor scenes covering 26 indoor scene types, with a total of 35,064 instances of objects labeled as 11 semantic categories (ceiling, floor, wall, window, chair, bed, sofa, table, televisions, furniture, objects), and was sampled in the experiments with 795 and 654 samples in the training and validation sets, respectively. The OccScanNet dataset was built based on the large-scale ScanNet[4] dataset to provide a richer and more scalable indoor scene than NYUv2, which focuses on 3D indoor occupancy of the scenes. It contains 45,755 training samples and 19,764 validation samples, which is 40 times more than NYUv2, and contains 11 semantic categories as NYUv2. In addition, OccScanNet_mini is a mini version of the OccScanNet dataset with 10% of the sample size of the OccScanNet dataset, containing 4639 training samples and 2007 validation samples.

Metric. In order to fully validate the effectiveness of our approach, we used the Intersection over Union of voxels (IoU) to evaluate the effect of semantic scene completion, the IoUs of each category of semantic scene completion to evaluate the prediction effect of each category, and the mean of the IoUs of each category of semantic scene completion (mIoU) to evaluate the overall performance of the model's semantic scene completion.

Experimental Details. We performed experiments on the NYUv2 dataset with an experimental epoch of 30 and a learning rate of 1e-4. Comparison experiments with the SOTA model on the Occscannet dataset were performed with an experimental epoch of 10 and a learning rate of 1e-4. All experiments were performed on a set of high performance GPUs equipped with Intel(R) Xeon(R) CPUs, 128 GB RAM, and NVIDIA V100 GPUs on a Linux server.

3.2 Ablation study

In this section, in order to demonstrate the effectiveness of our proposed MCCM, DCRNet, and KSDA modules, we conducted ablation experiments, and the experimental results are shown in Table 1.

Table 1. Ablation experiment on the NYUv2 dataset.

Baseline	MCCM	DCRNet	KSDA	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	televisions	furniture	objects	mIoU
√				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
√	√			47.45	16.46	93.48	16.10	**16.64**	17.46	49.30	41.00	**19.24**	24.20	35.54	19.02	31.68
√		√		47.53	15.00	93.24	16.76	13.99	17.39	49.20	42.65	18.23	28.76	34.60	18.87	31.70
√			√	47.68	17.40	93.49	16.31	15.90	17.52	49.39	40.60	18.66	24.75	35.02	**19.48**	31.68
√	√	√		47.82	16.22	93.42	16.86	15.53	18.02	49.84	41.97	18.99	25.29	35.19	19.05	31.85
√	√		√	47.61	16.23	**93.54**	16.79	16.52	17.30	49.94	42.67	18.46	26.10	**35.75**	18.23	31.95
√		√	√	47.68	**17.57**	93.45	16.73	14.05	17.23	51.11	**42.71**	17.74	**30.39**	34.40	18.36	32.16
√	√	√	√	**47.94**	16.33	93.16	**17.20**	15.78	17.90	**51.32**	42.02	18.98	28.30	35.25	19.37	**32.33**

It can be observed that, compared to the baseline model, the model trained with the MCCM module demonstrates significant improvements in the categories of ceiling, window, table, furniture, and objects. This is because the MCCM module calibrates the output of the 2D encoder, preserving multi-scale important features extracted through visual feature extraction while suppressing irrelevant information. Through this module, the network incorporates more multi-scale critical feature information during the 2D decoding process. This enhancement is particularly pronounced for highly complex categories like objects, and also yields substantial improvements for categories with significant scale variations such as ceiling, window, table, and furniture. However, the reduced IoU for chair, bed, and televisions categories stems from the MCCM module's emphasis on strengthening the 2D decoder's ability to comprehend global structures.

While this addresses the challenge of accurately completing objects with varying scales, it inadvertently diminishes the model's perception capability for categories with scale similarities like chair, bed, and televisions, thereby reducing their prediction effectiveness. However, the model achieves notable overall improvements in both IoU and mIoU, effectively validating the efficacy of the proposed module.

In addition, after integrating the DCRNet module into the baseline model, the overall performance of the model is improved. Analyzing the IoU of individual categories, significant enhancements are observed in semantic scene completion for ceiling, wall, sofa, televisions, furniture, and objects categories, while performance declines for window, chair, and bed classes. This improvement comes from our module’s recalibration of depth information, which produces a more refined depth distribution, thereby boosting the model’s holistic capabilities. In particular, substantial gains are achieved for categories with smoother surfaces, such as ceiling and wall, as well as highly complex ones, such as objects. However, for indoor furniture categories such as window, chair, and bed, the model tends to misclassify them as similar categories such as sofa, televisions, or furniture. This confusion reduces the IoU for the window, chair, and bed while increasing the IoU for their similar counterparts. However, the overall IoU and mIoU of the model increase, which validates the effectiveness of the proposed module.

Third, after incorporating the KSDA module into the baseline model, we observe exceptionally significant improvements in the categories of ceiling, wall, window, furniture, and objects. This is because our module effectively captures long-range dependencies while reducing information loss during downsampling, thereby alleviating the challenges of scene completion caused by complex scene scales and significant variations in object sizes during the semantic scene completion process. Through this module, the semantic information in the 3D encoder becomes more enriched, leading to remarkable enhancements for complex categories like furniture and objects, which rely heavily on detailed semantic understanding. However, in the categories of chair, bed, and televisions, performance degrades, due to the fact that the network lacks semantic information for small volume targets. However, the overall IoU and mIoU of the model show substantial improvements, sufficiently validating the efficacy of the proposed module.

Finally, each of our modules is not isolated, and each of them has its own focus. It can be seen that after adding the proposed MCCM, DCRNet, and KSDA modules at the same time, for the categories of chair, bed, televisions, and window, which are difficult to be improved by a single module, the new model has a great improvement in these categories. In conclusion, our method is targeted at occupying the prediction task in monocular indoor occupancy, and we have achieved good results in the experiments, verifying the effectiveness of the method.

3.3 Comparison and analysis with SOTA models

Performance on the NYUv2 dataset. We first conduct a performance comparison experiment with the SOTA model on the NYUv2 dataset, and the results obtained are shown in Table 2. As we can see, MDKOcc outperforms the SOTA model ISO in the evaluation metrics for categories including ceiling, wall, window, bed, sofa, table, televisions, furniture, and objects. Although MDKOcc achieves lower scores than the NDC-Scene model on the floor category and underperforms ISO in the chair category, it surpasses ISO overall when considering all evaluation metrics. Notably, MDKOcc achieves an IoU of 47.94% and an mIoU of 32.33%, exceeding ISO by 0.83% and 1.08%, respectively, which validates the superiority of MDKOcc.

In addition, we performed a visual analysis of ISO and MDKOcc, with results illustrated in Fig.5. Comparing the visualizations with ground truth, MDKOcc exhibits richer detail, smoother and more natural object surfaces, and more accurate category identification. This indicates that MDKOcc achieves finer detail processing in semantic scene completion, stronger global contextual relationships, and more precise semantic information, further validating its superiority.

Table 2. Performance comparison of SOTA models on the NYUv2 dataset.

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	televisions	furniture	objects	mIoU
LMSCNet^{rgb}[16]	Input^{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^{rgb}[10]	Input^{rgb} , Input^{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^{rgb}[3]	Input^{rgb} , Input^{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[2]	Input^{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[21]	Input^{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[22]	Input^{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
MDKOcc	Input^{rgb}	**47.94**	**16.33**	93.16	**17.20**	**15.78**	17.90	**51.32**	**42.02**	**18.98**	**28.30**	**35.25**	**19.37**	**32.33**

Performance on the OccScanNet dataset. We also conducted performance comparison experiments with the SOTA model on the OccScanNet dataset and the results obtained are shown in Table 3. It can be seen that MDKOcc outperforms the ISO model in terms of evaluation metrics on the categories of wall, window, bed, sofa, table, televisions, furniture, and objects. However, it is lower than the ISO model on the ceiling category and lower than the Monoscene model on the floor category. Although MDKOcc is lower than other methods in some of the categories, overall MDKOcc outperforms the SOTA model, as reflected in its IoU and mIoU, which reach 42.81% and 29.59%, respectively, which proves that MDKOcc is still superior on large datasets and confirms that MDKOcc has excellent generalisation performance.

In addition, we have also visualised ISO and MDKOcc separately, and the results are shown in Fig.6. It can be seen that in the four scenarios, the semantic scene complementation of MDKOcc is better than that of the ISO model, which further confirms that the MDKOcc model has excellent generalisation.

Table 3. Performance comparison of SOTA models on the OccScanNet dataset.

Method	Input	IoU	ceiling	floor	wall	window	chair	bed	sofa	table	televisions	furniture	objects	mIoU
Monoscene[2]	Input^{rgb}	41.16	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[22]	Input^{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
MDKOcc	Input^{rgb}	**42.81**	19.23	42.16	**23.04**	**19.52**	28.96	**43.59**	**43.20**	**30.13**	**13.14**	**36.42**	**26.07**	**29.59**

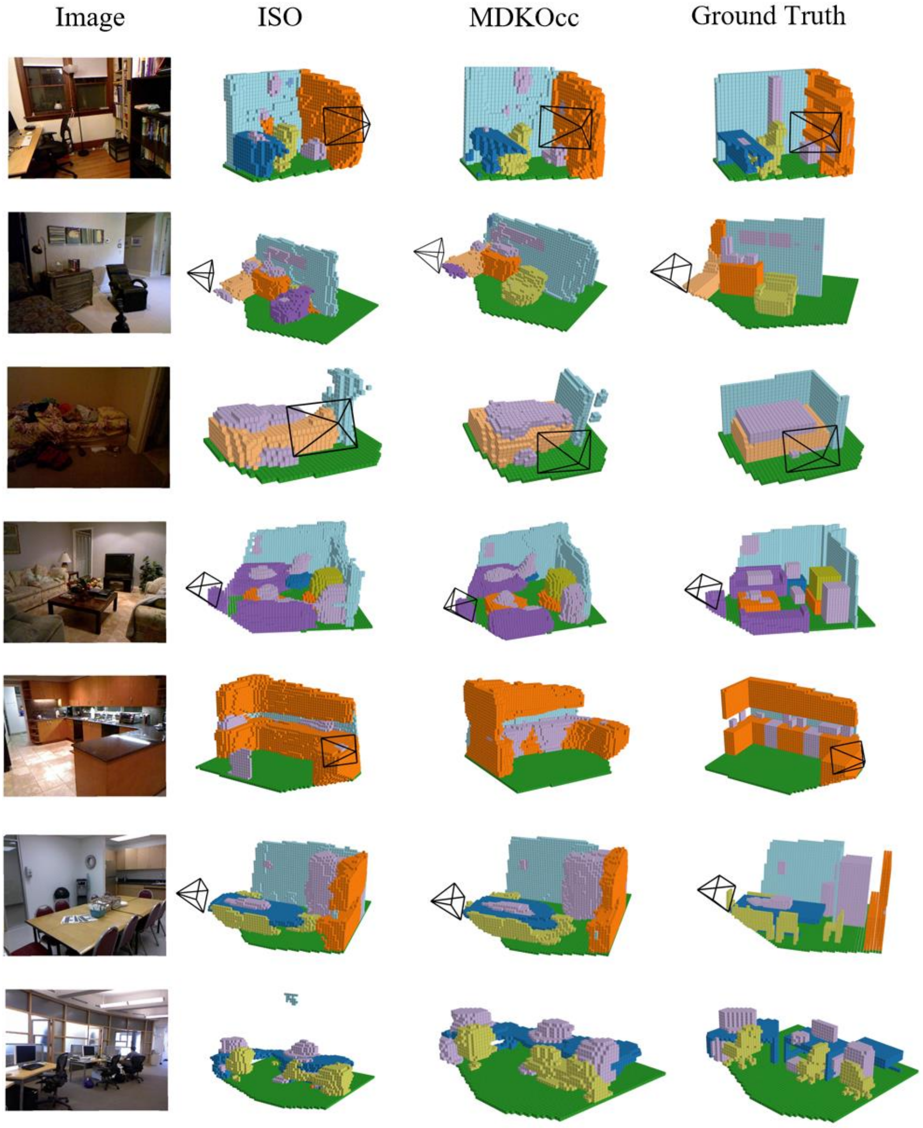

Fig. 5. Visual comparison of SOTA models on the NYUv2 dataset.

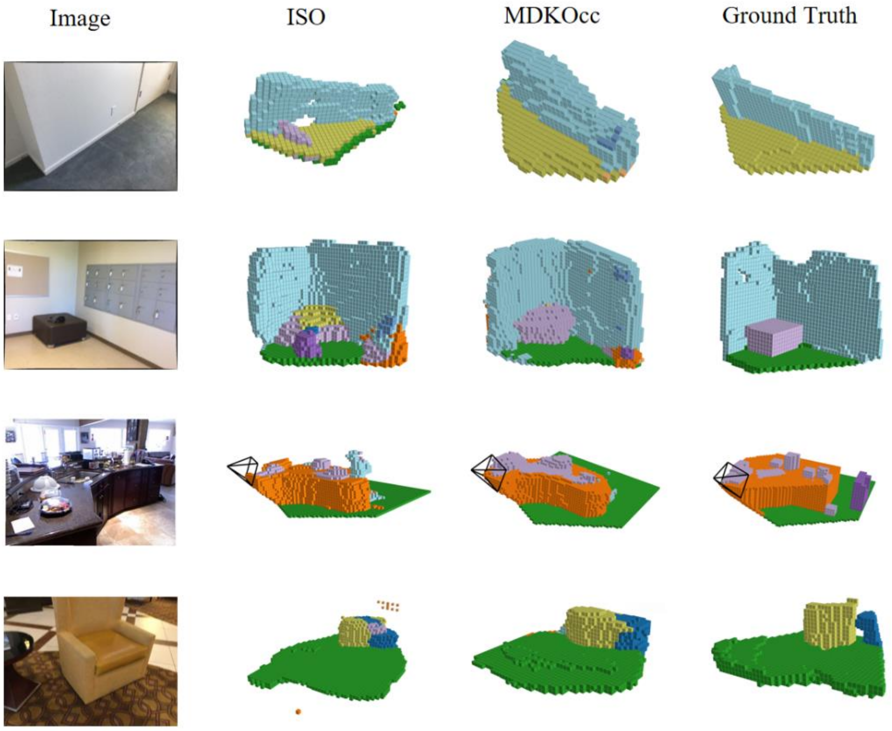

Fig. 6. Visual comparison of SOTA models on the OccScanNet dataset.

4 Discussion

4.1 Generalisability with limited data

In order to test the generalisation performance of MDKOcc on a finite dataset, we conducted experiments on the OccScanNet_mini dataset, and the experimental results are shown in Table 4. It can be seen that the IoU and mIoU of MDKOcc are still higher than that of ISO, which verifies that MDKOcc still has excellent generalisation on a finite dataset.

Table 4. Performance comparison on the OccScanNet_mini dataset.

Method	Input	IoU	mIoU
ISO[22]	Input^{rgb}	51.03	39.08
MDKOcc	Input^{rgb}	**51.56**	**39.43**

4.2 Limitations and future work

Although the MDKOcc model is rich in semantic information and reinforces the long-range dependency, which effectively mitigates the depth ambiguity problem, the effect on the OccScanNet dataset, which contains multi-view and multi-scale scenes, can be improved. This may be due to the fact that the scene categories of OccScanNet are complex and cover a wider range, and MDKOcc is more weak in learning these complex scenes, resulting in poor results. In the future, we will pay more attention to the situation in the real world and conduct research on more complex and extensive scenes to enhance the understanding of semantic information and better complement the semantic scenes.

5 Conclusion

In this paper, we propose a novel approach for mitigating the problem of semantic scene depth ambiguity in an indoor monocular occupancy prediction task. Our proposed MDKOcc method enhances the ability to complement the semantic scene. Ablation experiments confirm the effectiveness of our proposed Multi-scale Context Calibration Module, Depth Calibrated Residual Network and Kernel-split Depthwise Attention. In addition, comparison and analysis experiments with the SOTA model on the NYUv2 dataset confirm the superiority of our approach, as well as comparison and analysis experiments with the SOTA model on the OccScanNet dataset validate the generalisation of our approach. Finally, the comparison and analysis with the SOTA model on the OccScanNet_mini dataset verifies the excellent generalisation of our method even with limited data. In the future, we will focus on understanding semantic information and enhancing the complementation ability in complex semantic scenarios.

Acknowledgments. This research was funded by the National Natural Science Foundation of China (No. 62403076), the Humanities and Social Science Fund of Ministry of Education (No. 24YJCZH416) and Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province (New energy intelligent vehicle technology, 2024RC1029).

References

1. Cai, Y., Chen, X., Zhang, C., Lin, K.Y., Wang, X., Li, H.: Semantic scene completion via integrating instances and scene in-the-loop. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 324– 333 (2021)
2. Cao, A.Q., De Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022)

3. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4193–4202 (2020)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
5. Du, R., Feng, R., Gao, K., Zhang, J., Liu, L.: Self-supervised point cloud prediction for autonomous driving. IEEE Transactions on Intelligent Transportation Systems (2024)
6. Feng, T., Wang, W., Ma, F., Yang, Y.: Lsk3dnet: Towards effective and efficient 3d perception with large sparse kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14916–14927 (2024)
7. Jang, H.K., Kim, J., Kweon, H., Yoon, K.J.: Talos: Enhancing semantic scene completion via test-time adaptation on the line of sight. arXiv preprint arXiv:2410.15674 (2024)
8. Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J.: Spherical transformer for lidar-based 3d recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17545–17555 (2023)
9. Lavreniuk, M., Bhat, S.F., Müller, M., Wonka, P.: Evp: Enhanced visual percep- tion using inverse multi-attentive feature refinement and regularized image-text alignment. arXiv preprint arXiv:2312.08548 (2023)
10. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3351–3359 (2020)
11. Li, M., Liu, Z., Li, G., Han, D.: A multi-granularity semantic extraction method for text classification. In: International Conference on Intelligent Computing. pp. 224–236. Springer (2024)
12. Liu, X., Xie, H., Zhang, S., Yao, H., Ji, R., Nie, L., Tao, D.: 2d semantic-guided semantic scene completion. International Journal of Computer Vision pp. 1–20 (2024)
13. Liu, Y., Chen, R., Li, X., Kong, L., Yang, Y., Xia, Z., Bai, Y., Zhu, X., Ma, Y., Li, Y., et al.: Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21662–21673 (2023)
14. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Transactions on Graphics (ToG) **38**(6), 1–15 (2019)
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
16. Roldao, L., De Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 2020 International Conference on 3D Vision (3DV). pp. 111–119. IEEE (2020)
17. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. pp. 746–760. Springer (2012)
18. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4840– 4851 (2024)
19. Xia, Z., Liu, Y., Li, X., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y.: Scpnet: Semantic scene completion on point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17642–17651 (2023)

20. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14475–14485 (2023)
21. 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: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9421–9431. IEEE Computer Society (2023)
22. Yu, H., Wang, Y., Chen, Y., Zhang, Z.: Monocular occupancy prediction for scal- able indoor scenes. In: European Conference on Computer Vision. pp. 38–54. Springer (2024)
23. Zeid, K.A., Yilmaz, K., de Geus, D., Hermans, A., Adrian, D., Linder, T., Leibe, B.: Dino in the room: Leveraging 2d foundation models for 3d segmentation. arXiv preprint arXiv:2503.18944 (2025).
24. Wang, T., Ong, Z. C., Khoo, S. Y., Siow, P. Y., Zhang, J., & Wang, T. An enhanced generative adversarial network for longer vibration time data generation under variable operating conditions for imbalanced bearing fault diagnosis. Engineering Applications of Artificial Intelligence, 151, 110760. (2025).
25. Zhang, J., Yang, W., Chen, Y., Ding, M., Huang, H., Wang, B., ... & Du, R. Fast object detection of anomaly photovoltaic (PV) cells using deep neural networks. Applied Energy, 372, 123759. (2024).
26. Duan, Y., Meng, L., Meng, Y., Zhu, J., Zhang, J., Zhang, J., & Liu, X. MFSA-Net: Semantic Segmentation With Camera-LiDAR Cross-Attention Fusion Based on Fast Neighbor Feature Aggregation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. (2024).
27. Gao, K., Li, X., Hu, L., Liu, X., Zhang, J., Du, R., & Li, Y. STMF-IE: A Spatial-Temporal Multi-Feature Fusion and Intention-Enlightened Decoding Model for Vehicle Trajectory Prediction. IEEE Transactions on Vehicular Technology. (2024).
28. Wu, J., Zhang, J., Zhu, J., Duan, Y., Fang, Y., Zhu, J., ... & Meng, Y. Multi-scale convolution and dynamic task interaction detection head for efficient lightweight plum detection. Food and Bioproducts Processing, 149, 353-367. (2025).