MR-Occ: Efficient Camera-LiDAR 3D Semantic Occupancy Prediction Using Hierarchical Multi-Resolution Voxel Representation

Minjae Seong[1*], Jisong Kim[2*], Geonho Bang[1*], Hawook Jeong[3], and Jun Won Choi[4]

arXiv:2412.20480v1 [cs.CV] 29 Dec 2024

Abstract—Accurate 3D perception is essential for understanding the environment in autonomous driving. Recent advancements in 3D semantic occupancy prediction have leveraged camera-LiDAR fusion to improve robustness and accuracy. However, current methods allocate computational resources uniformly across all voxels, leading to inefficiency, and they also fail to adequately address occlusions, resulting in reduced accuracy in challenging scenarios. We propose MR-Occ, a novel approach for camera-LiDAR fusion-based 3D semantic occupancy prediction, addressing these challenges through three key components: Hierarchical Voxel Feature Refinement (HVFR), Multi-scale Occupancy Decoder (MOD), and Pixel to Voxel Fusion Network (PVF-Net). HVFR improves performance by enhancing features for critical voxels, reducing computational cost. MOD introduces an 'occluded' class to better handle regions obscured from sensor view, improving accuracy. PVF-Net leverages densified LiDAR features to effectively fuse camera and LiDAR data through a deformable attention mechanism. Extensive experiments demonstrate that MR-Occ achieves state-of-the-art performance on the nuScenes-Occupancy dataset, surpassing previous approaches by $+5.2\%$ in IoU and $+5.3\%$ in mIoU while using fewer parameters and FLOPs. Moreover, MR-Occ demonstrates superior performance on the SemanticKITTI dataset, further validating its effectiveness and generalizability across diverse 3D semantic occupancy benchmarks.

Index Terms—Autonomous driving, 3D Semantic Occupancy Prediction, Sensor Fusion.

I. INTRODUCTION

3D semantic occupancy prediction task aims to predict the occupancy and semantic information of fine-grained voxels surrounding the ego vehicle. This task provides a comprehensive volumetric scene representation that can be utilized by the path planner to enhance the safety of autonomous driving. While early research primarily focused on unimodal approaches using either LiDAR or camera data, recent methods have used multi-modal sensors, such as LiDAR and camera, to achieve more robust and accurate predictions.
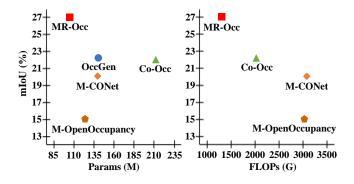

Fig. 1. **Accuracy vs Efficiency (Params/FLOPs) on nuScenes-Occupancy validation set.** MR-Occ achieves state-of-the-art performance with less computational cost than previous methods.

The key challenge in camera-LiDAR fusion for 3D semantic occupancy prediction lies in constructing robust 3D representations that effectively handle the distinct characteristics of each modality. LiDAR provides accurate 3D positional information, enabling accurate occupancy prediction, while cameras offer rich semantic details for object classification. Previous studies [1]–[3] have focused on integrating these two modalities by using 3D voxel representations generated from LiDAR data. To fuse camera features with 3D LiDAR features, these approaches predict depth from images, transform them from 2D to 3D views to create voxel-based image features, and then align these features with the LiDAR data in the voxel space. Finally, the fused features are processed using 3D convolution-based modules.

While these existing approaches have yielded great results, they encounter two main challenges. First, computational resources are uniformly allocated across all voxel representations, which is inefficient. In outdoor scenarios, only a small fraction of voxels are non-empty—approximately 2.47% according to the nuScenes-Occupancy dataset [1]. Therefore, considering the distribution of data across 3D voxels is essential for achieving more efficient occupancy state predictions.

Second, existing methods have underestimated the challenges posed by occlusion in real-world scenarios, leading to notable prediction inaccuracies in regions obscured from sensor view. These methods tend to either overlook hidden areas or assume equal visibility across all voxels. Thus, distinguishing between visible and non-visible areas is essential for achieving robust

[1]M. Seong and G. Bang are with the Department of Artificial Intelligence, Hanyang University, 04753 Seoul, Republic of Korea. (e-mail: mjseong@spa.hanyang.ac.kr, ghbang@spa.hanyang.ac.kr)

[2]J. Kim is with the Department of Electrical Engineering, Hanyang University, 04753 Seoul, Republic of Korea. (e-mail: jskim@spa.hanyang.ac.kr)

[3]H. Jeong is with the RideFlux Inc., 07217 Seoul, Republic of Korea. (e-mail: hawook@rideflux.com)

[4]J. W. Choi is with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, 08826, Korea. (e-mail: junwchoi@snu.ac.kr)(*Corresponding author: Jun Won Choi*)

*denotes equal contribution.

and accurate model predictions in complex environments.

In this paper, we present MR-Occ, a novel approach for camera-LiDAR fusion-based semantic occupancy prediction. The key ideas of MR-Occ are three-fold.

First, we introduce Hierarchical Voxel Feature Refinement (HVFR) method, which selectively increases the resolution of voxels for efficient voxel encoding. Beginning with low resolution voxels, HVFR identifies core voxels, that captures important regions in a scene, based on occupancy confidence scores. By subdividing these critical voxels into smaller ones at finer resolutions, HVFR can capture important details and more refined voxel representation while maintaining low computational cost.

Second, we found that regions not visible to sensors due to occlusion by objects pose significant challenges for 3D occupancy prediction models. These models may need to rely on prior object knowledge or temporal information to determine occupancy in these ambiguous regions. When predictions for these regions are treated equally with those for non-ambiguous areas, the models may struggle to accurately predict occupancy. To address this issue, we divide Non-empty class into Occluded and Non-occluded classes, where Occluded class is assigned to voxels that are not visible to sensors but are labeled as Non-empty. By incorporating the Occluded class, we can train the model to classify a new class while treating the Occluded class as Non-empty class in the inference phase. This approach effectively reduces the burden on the models and improves occupancy prediction accuracy. Importantly, it does not require additional computational resources, as only a simple logic for labeling the ambiguity state is needed.

Finally, we introduce the Pixel to Voxel Fusion Network (PVF-Net), an efficient feature fusion strategy that associates camera features with each voxel using a deformable attention mechanism guided by voxel queries generated from densified LiDAR features. Unlike most previous methods, which suffer from misalignment between camera and LiDAR features due to inaccurate depth estimation, our approach does not rely on depth prediction for feature fusion. This effectively resolves the misalignment issue, leading to significant performance gains.

MR-Occ achieves state-of-the-art performance on the nuScenes-Occupancy dataset for camera-LiDAR-based 3D semantic occupancy prediction task. It demonstrates remarkable performance gains of $+5.2\%$ in IoU and $+5.3\%$ in mIoU over the previous best method, OccGen [3] while requiring fewer parameters and FLOPs, as shown in Figure 1.

The key contributions of our work are as follows:

- We propose MR-Occ, a new 3D semantic occupancy prediction model that effectively fuses LiDAR and camera features.
- We propose a Hierarchical Voxel Feature Refinement method that successively increases the resolution of core voxels that capture key scene information. Our approach significantly enhances the efficiency of voxel encoding.
- We propose a Multi-scale Occupancy Decoder that introduces a new Occluded state for voxels in the occupancy prediction problem. Our model is assigned a task of classifying Occluded state along with existing occupancy classes. This approach enables the model can effectively

distinguish between occluded and non-occluded areas, thereby improving overall occupancy prediction accuracy.
- We propose a Pixel to Voxel Fusion Network that aligns 2D camera features with densified LiDAR features and fuses them with adaptive weights using deformable cross-attention.
- The code will be publicly available.

II. RELATED WORK

A. Unimodal 3D Semantic Occupancy Prediction

3D Semantic Occupancy Prediction (SOP) has emerged as a pivotal task in autonomous driving, leading to diverse research on LiDAR and camera-based approaches. Early research in 3D SOP was predominantly focused on LiDAR-centric approaches [4]–[6] based on the SemanticKITTI dataset [7]. Recent advancements have expanded these approaches by transferring rich semantic information from multi-frame LiDAR models [8], leveraging generative models [3], and constructing 2D tri-perspective view representations [9].

In parallel, camera-based methods have garnered attention due to their cost-effectiveness and capacity to provide detailed visual information. A notable example is MonoScene [10], which pioneered effective 3D SOP using only a single RGB image, thereby opening new research avenues. Recent camera-based approaches are primarily categorized into two paradigms: forward projection and backward projection. Forward projection explicitly maps 2D image features into 3D space by estimating depth and applying geometric transformations. OccFormer [11] advances this approach by introducing a dual-path transformer that captures long-range dependencies and effectively handles dynamic 3D features. To optimize computational efficiency, methods such as [12]–[14] combine 2D convolutions with channel-to-height transformations and utilize sparse 3D representations. In contrast, backward projection implicitly learns complex relationships between 2D and 3D representations, transforming 2D image features into 3D volumetric features through transformer architectures. Within this framework, TPV-Former [15] and VoxFormer [16] employ tri-perspective views and depth-estimated voxel queries, respectively, to mitigate the computational cost associated with attention mechanisms.

B. Multimodal 3D Semantic Occupancy Prediction

While unimodal approaches have progressed significantly in 3D SOP, integrating multi-sensor data is crucial for comprehensive environmental understanding. Extensive research in 3D object detection, a closely related task to 3D SOP, has explored various sensor fusion strategies. Some approaches, inspired by multi-view 3D object detection, employ explicit view transformations [17] to align camera and LiDAR data within a unified Bird's Eye View (BEV) [18], [19]. Others utilize attention-based methods [20], [21] with transformer architectures that use object and BEV queries to dynamically focus on relevant features from multiple sensor modalities.

While various sensor fusion techniques have been explored in 3D object detection task, multi-modal 3D SOP primarily utilizes explicit view transform methods to convert camera
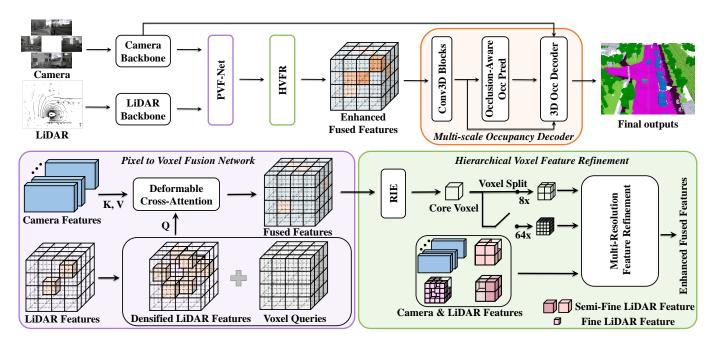
Fig. 2. Overall architecture of MR-Occ: Camera and LiDAR features are extracted from modality-specific backbone networks. The PVF-Net densifies LiDAR features and adaptively fuses them with image features using a deformable cross-attention mechanism. The HVFR module uses Resolution Importance Estimator (RIE) to identify core voxels, and then enhances the fused features through Multi-Resolution Feature Refinement using these core voxels. Finally, Multi-scale Occupancy Decoder (MOD) predicts an 'occluded' class for occluded areas and performs fine-grained occupancy prediction.

data into 3D voxel features. These camera 3D voxel features are then fused with LiDAR features in the 3D voxel space. CONet [1] employs adaptive fusion to integrate LiDAR and camera information, coupled with a coarse-to-fine prediction strategy. Similarly, Co-Occ [2] introduces a Geometric- and Semantic-aware Fusion (GSFusion) module, enhancing LiDAR features by incorporating neighboring camera features through a K-nearest neighbors (KNN) search. These enriched LiDAR features are then concatenated with the original LiDAR and camera voxel features. In contrast, OccGen [3] tackles the misalignment between camera and LiDAR features by applying a hard 2D-to-3D view transformation and utilizing a geometry mask.

Despite these advancements, existing multimodal approaches generate 3D voxel features separately for each modality, increasing computational costs during the fusion process. Additionally, by assigning equal importance to all voxels, these methods lead to unnecessary computations and inefficient use of resources, ultimately limiting performance improvements.

III. MR-Occ

The overall architecture of MR-Occ is illustrated in Figure 2. Our model is a camera-LiDAR fusion framework for 3D Semantic Occupancy Prediction that identifies core voxels and intensively enhances their features. In Section 3.1, we introduce the Pixel to Voxel Fusion Network (PVF-Net) method to fuse LiDAR voxel features with multi-camera features. Then, we present the Hierarchical Voxel Feature Refinement (HVFR) module for identifying core voxels and enhancing their features in Section 3.2. Finally, the Multi-scale Occupancy Decoder (MOD) for performing multi-scale 3D semantic occupancy prediction is described in Section 3.3.

A. Pixel to Voxel Fusion Network

Prior to integrating LiDAR and image features, we extract both features using separate backbone networks. For LiDAR data, we employ 3D sparse convolutional layers to compute multi-resolution voxel features $F_L = \{F_L^1, F_L^2, F_L^4\}$, where $F_L^i \in \mathbb{R}^{C_L^i \times H_L^i \times W_L^i \times D_L^i}$. Here, i represents the downsampling scale, (H_L^i, W_L^i, D_L^i) correspond to the 3D spatial dimensions, and C_L^i represents the number of channels. For camera data, we utilize a ResNet-50 backbone integrated with a Feature Pyramid Network (FPN) to derive multi-view features $F_I \in \mathbb{R}^{N \times H_I \times W_I \times C_I}$, where N is the number of cameras, (H_I, W_I) are the 2D spatial dimensions of the feature maps, and C_I denotes the channel dimension.

Existing camera-LiDAR methods [1]–[3] transform image features F_I to 3D voxel representation to fuse with LiDAR features. However, this method can lead to positional misalignment as the 3D features derived from camera are inherently inaccurate. To address this issue, we introduce the Pixel-to-Voxel Fusion Network (PVF-Net), which enhances multi-modal fusion through densified LiDAR features. PVF-Net expands the receptive field around non-empty voxels, enabling LiDAR features to guide the seamless integration of 2D image features into 3D voxel representation.

First, we downsample LiDAR feature F_L^4 using 3D sparse convolutions to obtain F_L^8 and F_L^{16}. Let the non-empty voxel features of F_L^i be denoted as $V_L^i \in \mathbb{R}^{C_L^i \times N^i}$, where N^i represents the number of non-empty voxels, and their corresponding indices as $G^i \in \mathbb{Z}^{3 \times N^i}$ at each scale i. We then concatenate these features across different scales

$$\tilde{V}_L^4 = [V_L^4, V_L^8, V_L^{16}], \quad \tilde{G}_L^4 = [G^4, 2 \cdot G^8, 4 \cdot G^{16}], \quad (1)$$

Semi-Fine Resolution Refinement

Fig. 3. Multi-Resolution Feature Refinement module. The subdivided core voxels combine features sampled from the same resolution LiDAR features and camera features to capture fine-grained details. The multi-resolution features are fused based on a 3D sparse convolution.

where $[\cdot]$ denotes the concatenation operation. Note that the factors 2 and 4 in $2 \cdot G^8$ and $4 \cdot G^{16}$ align the scales of G^8 and G^{16} with G^4. The final dense feature map $\tilde{F}_L^4$ is computed by averaging the overlapping non-empty features at each voxel location

$$\tilde{F}_L^4(x, y, z) = \frac{1}{|S(x, y, z)|} \sum_{k \in S(x,y,z)} \tilde{V}_L^4(k), \quad (2)$$

where $S(x, y, z)$ denotes the set of indices k for which the corresponding voxel in $\tilde{G}_L^4$ maps to the voxel (x, y, z).

Next, we fuse $\tilde{F}_L^4$ with the image features F_I using deformable cross-attention. In this process, 3D voxel queries are projected onto the image plane, where nearby pixel features serve as keys and values. To guide the alignment of 2D image feature within the 3D voxel space, we update the random queries Q_v by adding the densified LiDAR features $\tilde{F}_L^4$. The LiDAR guided query Q_v' is then formulated as

$$Q_v' = \tilde{F}_L^4 + Q_v. \quad (3)$$

The Pixel-to-Voxel fusion is then performed as

$$F_M^4 = \frac{1}{|V_{\text{hit}}|} \sum_{i \in V_{\text{hit}}} \sum_{j=1}^{N_{\text{ref}}} \text{DA}(Q_v'(p), P(p, i, j), F_I^i), \quad (4)$$

where DA represents the deformable attention function, $Q_v'(p)$ denotes the LiDAR guided query at position p, and $P(p, i, j)$ projects it to the j-th reference point on the i-th camera view. F_I^i represents features from the i-th camera, N_{ref} is the number of reference points per query, and V_{hit} is the set of cameras where the projected point is visible.

B. Hierarchical Voxel Feature Refinement

Existing approaches often rely on downsampled fused features that are coarser than the ground truth voxel resolution, which hinders the accurate prediction of small objects and fine boundary details. While using finer-scale features could mitigate these issues, it would come at the cost of significantly increased computational complexity. To address this, we introduce a Hierarchical Voxel Feature Refinement (HVFR) module that adaptively refines the critical voxels within the feature map.

We first apply a Resolution Importance Estimator (RIE) to the fused feature map to determine the necessary level of detail for each voxel. This process yields a voxel-wise importance map R as given by the equation

$$R = \sigma(\text{Conv}_{3D}(F_M^4)), \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\text{Conv}_{3D}(\cdot)$ refers to a 3D convolutional layer. Voxels are selectively refined based on the value of $R(x, y, z)$ at each coordinate (x, y, z) in comparison to the predefined thresholds τ_1 and τ_2. Specifically, if $R(x, y, z) \geq \tau_1$, the voxel is assigned to the Semi-fine Resolution set $\mathcal{S}$, and if $R(x, y, z) \geq \tau_2$, it is further included in the Fine Resolution set $\mathcal{F}$.

For voxels in the Semi-fine Resolution set $\mathcal{S}$, each voxel is uniformly subdivided into eight smaller sub-voxels. The corresponding LiDAR features are extracted from a finer resolution LiDAR feature map F_L^2, and the image features F_I^2 for these sub-voxels are obtained as follows

$$F_I^2(x_i', y_i', z_i') = \text{Proj}(F_I, (x_i', y_i', z_i')), \quad i = 1, \ldots, 8, \quad (6)$$

where $\text{Proj}(\cdot)$ represents the operation of projecting the 3D positions (x_i', y_i', z_i') onto the image plane and retrieving the corresponding image features from F_I. These two features are then concatenated channel-wise and passed through a 1×1 convolutional layer to generate a refined feature F_S^2 for the Semi-fine Resolution set, as follows

$$F_S^2 = \text{Conv}_{1 \times 1}\left([F_L^2, F_I^2]\right). \quad (7)$$

For the Fine Resolution set $\mathcal{F}$, we apply a more detailed refinement process. Each voxel is uniformly subdivided into 64 finer sub-voxels, with the corresponding LiDAR features extracted from the finest resolution feature map F_L^1 and the image features obtained as follows

$$F_I^1(x_j'', y_j'', z_j'') = \text{Proj}(F_I, (x_j'', y_j'', z_j'')), \quad j = 1, \ldots, 64. \quad (8)$$

The refined feature F_F^1 is then given by

$$F_{\mathcal{F}}^1 = \text{Conv}_{1 \times 1}\left([F_L^1, F_I^1]\right). \quad (9)$$

Finally, we apply multi-scale feature fusion to the hierarchical voxel features F_S^2 and F_F^1 to obtain selectively refined features for core voxels. This fusion process is formulated as follows

$$F_E^4 = \text{SConv}_2\left(\text{SConv}_1\left(F_{\mathcal{F}}^1\right) + F_S^2\right) + F_M^4, \quad (10)$$

where SConv_1 and SConv_2 represent 3D sparse convolutions applied sequentially to integrate the refined features with the original fused feature map, F_M^4. The resulting feature map F_E^4 offers a more comprehensive scene representation, effectively synthesizing selectively refined details with broader contextual information.

C. Multi-scale Occupancy Decoder

3D semantic occupancy prediction requires dense predictions across the entire 3D space, including both visible and occluded voxels. However, previous studies have often overlooked visibility considerations in their occupancy state prediction frameworks. This oversight may limit the model's ability to

fully understand the scene, potentially reducing prediction accuracy. To overcome this challenge, we introduce the Occlusion-aware Occupancy Prediction (OOP) module. This module classifies each voxel grid as empty, non-occluded, or occluded, thereby improving the model's robustness and overall performance.

We extend conventional voxel ground truth (GT) by integrating semantic classes with three additional labels: 'non-occluded', 'occluded', and 'empty'. To assign these labels, we employ a ray-casting process with both LiDAR and camera data. Voxels containing LiDAR points or corresponding to projected image pixels are labeled 'non-occluded' voxels, subsequent voxels that have already been assigned a class label are marked as 'occluded'. The remaining voxels are labeled 'empty'. The final label is determined by combining results from both modalities: a voxel is labeled 'non-occluded' if either modality identifies it as such, 'occluded' if both modalities agree, and 'empty' if either modality classifies it as empty.

The enhanced fused features F_E generated by the HVFR module are first processed through Conv3D blocks. These processed features are then fed into the Occlusion-Aware Occupancy Prediction module, which consists of a sequence of Conv3D-BN-ReLU-Conv3D layers. This module outputs $O^4 \in \mathbb{R}^{D_L^4 \times H_L^4 \times W_L^4 \times 21}$, where 18 of the 21 channels correspond to semantic occupancy classes, and the remaining 3 represent occlusion-aware classes (empty, non-occluded, or occluded). Finally, the 3D Occupancy Decoder, which is equivalent to the occupancy head in M-CONet, processes the voxels predicted as non-occluded or occluded in O^4. This decoder then predicts the final fine-grained semantic occupancy $O^1 \in \mathbb{R}^{D_L^1 \times H_L^1 \times W_L^1 \times 21}$, providing a detailed semantic classification for each relevant voxel in the scene.

D. Loss Function

Our model is optimized using a comprehensive loss function, following the method presented in CONet [1]. The overall loss, $\mathcal{L}_{\text{total}}$, is formulated as follows

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{ls}} + \mathcal{L}_{\text{geo\_scal}} + \mathcal{L}_{\text{sem\_scal}} + \mathcal{L}_{\text{rs}} + \mathcal{L}_{\text{ocl}} \quad (11)$$

where $\mathcal{L}_{\text{ce}}$ represents the cross-entropy loss, and $\mathcal{L}_{\text{ls}}$ denotes the Lovász-Softmax loss [22], both of which are essential for semantic occupancy prediction. The affinity losses, $\mathcal{L}_{\text{geo\_scal}}$ and $\mathcal{L}_{\text{sem\_scal}}$ [10], are incorporated to enhance scene-wise and class-wise metrics. Additionally, $\mathcal{L}_{\text{rs}}$ is the binary cross-entropy loss employed for the resolution importance estimator, which is crucial for identifying key voxels. The ground truth for the resolution importance estimator is determined by assigning binary labels to each voxel, indicating whether it is occupied. Finally, $\mathcal{L}_{\text{ocl}}$ is an occlusion-aware loss based on cross-entropy, designed for the Occlusion-aware Occupancy Prediction module to address visibility constraints from input sensors. Each term in the loss function contributes to improving the overall performance of the model in 3D semantic occupancy prediction.

IV. EXPERIMENTS

A. Datasets

We evaluate MR-Occ on two large-scale datasets: nuScenes-Occupancy [1] and SemanticKITTI [7]. The nuScenes-Occupancy dataset provides dense semantic occupancy annotations for 1,000 scenes, with voxel grid annotations spanning $[-51.2\,\text{m}, 51.2\,\text{m}]$ in both X and Y directions, and $[-5\,\text{m}, 3\,\text{m}]$ in the Z direction, at a resolution of $512 \times 512 \times 40$. Each voxel is assigned one of 18 labels, comprising 17 semantic categories and 1 empty category. SemanticKITTI, based on the KITTI Odometry Benchmark [23], consists of 22 sequences containing LiDAR scans and front camera images. The annotations use a $256 \times 256 \times 32$ voxel grid, with voxel grid coordinates spanning $[0\,\text{m}, 51.2\,\text{m}]$ along the X-axis, $[-25.6\,\text{m}, 25.6\,\text{m}]$ along the Y-axis, and $[-2\,\text{m}, 4.4\,\text{m}]$ along the Z-axis. Each voxel is assigned to one of 21 classes (19 semantic, 1 free, and 1 unknown).

We evaluate the performance of our method using two widely adopted metrics: *Intersection over Union (IoU)* for geometric accuracy and *mean Intersection over Union (mIoU)* for semantic-aware perception quality.

a) Intersection over Union (IoU): IoU measures the voxel-level geometric accuracy as the ratio of the intersection volume to the union volume as

$$\text{IoU} = \frac{V_{\text{pred}} \cap V_{\text{gt}}}{V_{\text{pred}} \cup V_{\text{gt}}}, \quad (12)$$

where V_{pred} represents the predicted occupied voxels and V_{gt} represents the ground truth occupied voxels.

b) Mean Intersection over Union (mIoU): mIoU evaluates both occupancy prediction and its semantic consistency by averaging IoU across all C semantic classes as

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{V_{\text{pred},c} \cap V_{\text{gt},c}}{V_{\text{pred},c} \cup V_{\text{gt},c}}, \quad (13)$$

where $V_{\text{pred},c}$ and $V_{\text{gt},c}$ denote the predicted and ground truth voxels for class c, respectively.

B. Implementation Details

Our framework utilizes ResNet-50 [24] with a Feature Pyramid Network (FPN) [25] as the camera backbone and SECOND [26] as the LiDAR backbone. The input image resolution varies between datasets: for the nuScenes-Occupancy dataset, the images are set to 900×1600 pixels, while for the SemanticKITTI dataset, the images are resized to 384×1280 pixels to standardize the input dimensions. For LiDAR input, we apply 10 historical sweeps for the nuScenes-Occupancy dataset to improve temporal consistency, whereas for the SemanticKITTI dataset, we use the original LiDAR point clouds without any modifications.

The model is implemented using the MMDetection3D codebase [27]. We adopt a consistent training strategy across both datasets. Specifically, the model is trained for 15 epochs with a batch size of 1. Optimization is performed using the AdamW optimizer, with a weight decay of 0.01 and an initial learning rate of 3×10^{-4}. The learning rate is adjusted using a cosine learning rate scheduler with a linear warm-up phase applied during the first 500 iterations to stabilize the training process. To improve generalization, we employ data augmentation strategies inspired by BEVDet [28]. These include both Image Data Augmentation, such as random flipping, scaling, and photometric distortion, and BEV Data Augmentation, which

Method	Mod.	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	mammade	vegetation
C-OpenOccupancy [1]	C	19.3	10.3	9.9	6.8	11.2	11.5	6.3	8.4	8.6	4.3	4.2	9.9	22.0	15.8	14.1	13.5	7.3	10.2
C-CONet [1]		20.1	12.8	13.2	8.1	15.4	17.2	6.3	11.2	10.0	8.3	4.7	12.1	31.4	18.8	18.7	16.3	4.8	8.2
SparseOcc [14]		21.8	14.1	16.1	9.3	15.1	18.6	7.3	9.4	11.2	9.4	7.2	13.0	31.8	21.7	20.7	18.8	6.1	10.6
C-OccGen [3]		23.4	14.5	15.5	9.1	15.3	19.2	7.3	11.3	11.8	8.9	5.9	13.7	34.8	22.0	21.8	19.5	6.0	9.9
C-MR-Occ		25.6	16.2	17.3	9.9	16.8	21.2	8.2	12.7	12.9	10.1	7.5	14.3	38.9	25.3	24.7	20.6	8.0	11.2
L-OpenOccupancy [1]	L	30.8	11.7	12.2	4.2	11.0	12.2	8.3	4.4	8.7	4.0	8.4	10.3	23.5	16.0	14.9	15.7	15.0	17.9
L-CONet [1]		30.9	15.8	17.5	5.2	13.3	18.1	7.8	5.4	9.6	5.6	13.2	13.6	34.9	21.5	22.4	21.7	19.2	23.5
L-OccGen [3]		31.6	16.8	18.8	5.1	14.8	19.6	7.0	7.7	11.5	6.7	13.9	14.6	36.4	22.1	22.8	22.3	20.6	24.5
L-MR-Occ		**35.7**	24.1	28.6	13.6	22.1	29.0	13.5	20.4	26.4	16.1	18.3	23.2	**38.5**	25.1	**26.2**	25.7	28.6	30.3
M-OpenOccupancy [1]	M	29.1	15.1	14.3	12.0	15.2	14.9	13.7	15.0	13.1	9.0	10.0	14.5	23.2	17.5	16.1	17.2	15.3	19.5
M-CONet [1]		29.5	20.1	23.3	13.3	21.2	24.3	15.3	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
CO-Occ [2]		30.6	21.9	26.5	16.8	22.3	27.0	10.1	20.9	20.7	14.5	16.4	21.6	36.9	23.5	25.5	23.7	20.5	23.5
OccGen [3]		30.3	22.0	24.9	16.4	22.5	26.1	14.0	20.1	21.6	14.6	17.4	21.9	35.8	24.5	24.7	24.0	20.5	23.5
MR-Occ		35.5	27.3	**30.5**	**22.9**	**26.6**	**30.7**	**17.3**	**28.8**	**35.4**	**21.5**	**20.7**	**26.4**	38.1	**26.8**	25.9	**26.2**	**28.6**	30.4

TABLE I

3D SEMANTIC OCCUPANCY PREDICTION RESULTS ON NUSCENES-OCCUPANCY VALIDATION SET. THE MODEL IS TRAINED ON NUSCENES-OCCUPANCY TRAIN SET AND EVALUATED ON NUSCENES-OCCUPANCY VALIDATION SET. MOD.: MODALITY. C: CAMERA. L: LIDAR. M: CAMERA AND LIDAR. CONST. VEH.: CONSTRUCTION VEHICLE, DRIVE. SURF.: DRIVABLE SURFACE. THE BEST PERFORMANCE IS IN BOLDFACE.

Methods	PVF-Net	HVFR	OOP	IoU	mIoU
M-CONet				29.5	20.1
MR-Occ	✓			34.4	25.9
		✓		32.4	26.4
	✓		✓	34.5	26.2
	✓	✓		34.5	27.1
	✓	✓	✓	**35.5**	**27.3**

TABLE II

ABLATION STUDY ON THE NUSCENES-OCCUPANCY VALIDATION SET. PVF-NET: PIXEL TO VOXEL FUSION NETWORK. HVFR: HIERARCHICAL VOXEL FEATURE REFINEMENT. OOP: OCCLUSION-AWARE OCCUPANCY PREDICTION.

Methods	IoU	mIoU
M-CONet	29.5	20.1
M-CONet-DA	31.9	23.4
PVF-Net w/o LD	33.8	25.2
PVF-Net	**34.4**	**25.9**

TABLE III

ABLATION STUDY OF PVF-NET COMPONENTS. M-CONET-DA: M-CONET WITH 2D-TO-3D VIEW TRANSFORM REPLACED BY DEFORMABLE ATTENTION. LD: LIDAR FEATURES DENSIFICATION.

Methods	Params	GFLOPs	FPS	IoU	mIoU
M-OpenOccupancy	117M	3045	**4.0**	29.1	15.1
M-CONet	137M	3089	2.8	29.5	20.1
Co-Occ	205M	2028	2.4	30.6	21.9
OccGen	137M	-	2.3	30.3	22.0
MR-Occ	**106M**	**1334**	3.2	**35.5**	**27.3**

TABLE IV

ABLATION STUDY COMPARING EFFICIENCY AND ACCURACY. ALL MODELS EXCEPT OCCGEN WERE INFERRED ON AN NVIDIA RTX 3090 GPU.

consists of random rotation, scaling, and translation within the Bird's-Eye View space. These augmentations ensure that the model can robustly handle variations in both image and LiDAR inputs during training.

All experiments are conducted on a system equipped with four NVIDIA RTX 3090 GPUs and an Intel Xeon Silver 4210R CPU. The total training time is approximately 48 hours for the nuScenes-Occupancy dataset and 16 hours for the SemanticKITTI dataset.

The nuScenes-Occupancy dataset builds upon the publicly available nuScenes dataset [29], which can be accessed at https://www.nuscenes.org/. The semantic occupancy annotations are provided by the OpenOccupancy project and are available at https://github.com/JeffWang987/OpenOccupancy. Similarly, the SemanticKITTI dataset is an extension of the KITTI Odometry Benchmark [23], with voxel-wise semantic occupancy annotations publicly accessible at http://www.semantic-kitti.org/.

C. Main Results

nuScenes-Occuancy dataset. Table I presents a performance comparison on the nuScenes-Occupancy validation set. The proposed MR-Occ model achieves state-of-the-art results in Camera-only, LiDAR-only and multimodal configurations. The C-MR-Occ model, which applies HVFR and MOD to the C-CONet model, achieved 1.7% higher IoU and 1.8% higher mIoU compared to C-OccGen in the Camera-only configuration. This demonstrates that the proposed approach can effectively

CAM FRONT LEFT	Ground Truth	M-CONet	MR-Occ

barrier   **bicycle**   **bus**   **car**   **construction vehicle**   **motorcycle**   **pedestrian**   **traffic cone**

trailer   **truck**   **driveable surface**   **other flat**   **sidewalk**   **terrain**   **manmade**   **vegetation**
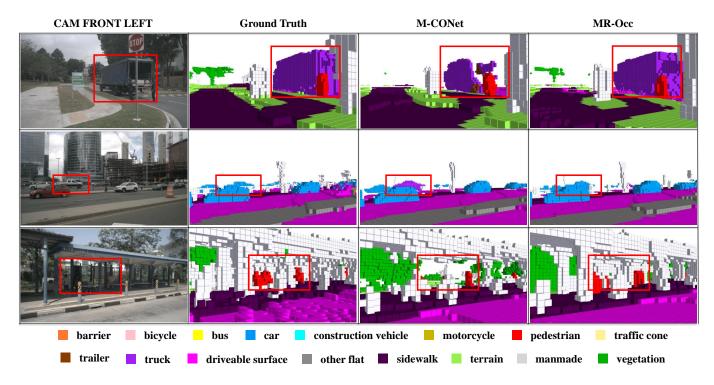
Fig. 4. Qualitative results comparing MR-Occ and M-CONet predictions: The red boxes highlight areas where MR-Occ shows improved accuracy in detecting objects, particularly in object boundary regions, occlusion scenarios and small objects.

Methods	Modality	IoU	mIoU
M-CONet		57.6	22.9
Co-Occ	C+L	57.5	22.4
MR-Occ		**58.4**	**25.6**

TABLE V
COMPARISON ON SEMANTICKITTI TEST SET. RESULTS FOR M-CONET
AND CO-OCC WERE REPRODUCED USING THEIR RESPECTIVE OFFICIAL
CODES.

produce reliable occupancy prediction results using only camera data. In the LiDAR-only scenario, where the deformable cross-attention module within PVF-Net is not utilized, the L-MR-Occ variant outperforms L-OccGen [3] by 3.9% in IoU and 7.3% in mIoU, highlighting the robustness of the proposed voxel-level fusion strategies. Within the multimodal setting, MR-Occ surpasses the baseline M-CONet [1] by 6.0% in IoU and 7.2% in mIoU. When compared to the state-of-the-art OccGen model, MR-Occ provides a 5.2% improvement in IoU and a 5.3% improvement in mIoU. These results underscore the model's ability to effectively integrate complementary sensor inputs and enhance spatial reasoning. Notably, while existing multimodal approaches often experience more than a 1% reduction in IoU due to sensor misalignment, MR-Occ restricts this degradation to only 0.2%, indicating effective mitigation of alignment challenges.

D. Ablation studies

Component analysis. Table II presents an ablation study demonstrating the contribution of each component in our proposed MR-Occ on the nuScenes-Occupancy validation set.

The Pixel to Voxel Fusion Network (PVF-Net), which employs deformable cross-attention guided by densified LiDAR features to fuse 2D camera features with 3D voxel features, significantly enhances performance. When applying the PVF-Net module to the baseline M-CONet, we observe significant performance improvements, with IoU increasing by 4.9% and mIoU by 5.8%. The Hierarchical Voxel Feature Refinement (HVFR) module, which focuses on core voxels and fuses fine-grained multimodal features, provides a significant boost when added to M-CONet, raising IoU by 2.9% and mIoU by 6.3%. Moreover, adding HVFR on top of the PVF-Net further increases IoU by 0.1% and mIoU by 1.2%. Finally, the integration of the Occlusion-aware Occupancy Prediction (OOP) module as an auxiliary task to predict occluded regions yields an additional 1.0% improvement in IoU and 0.2% gain in mIoU. The OOP module focused on predicting the occupancy of occluded voxels, which significantly improved IoU performance.

Effects of PVF-Net. Table III presents an ablation study on the nuScenes-Occupancy validation set, assessing the effectiveness of our PVF-Net components. Starting with M-CONet as the baseline, which achieves 29.5% IoU and 20.1% mIoU, we introduce M-CONet-DA, replacing the 2D-to-3D view transform with transformer-based feature extraction utilizing deformable attention and learnable voxel queries. This modification increases performance to 31.9% IoU and 23.4% mIoU. In the PVF-Net w/o LD configuration, LiDAR features are integrated into voxel queries without densification, further improving performance to 33.8% IoU and 25.2% mIoU, highlighting the importance of LiDAR's spatial information. Finally, the full PVF-Net, which includes LiDAR feature densification, achieves the best results with 34.4% IoU and

Method	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
M-CONet [1]	59.4	35.2	21.0	**2.0**	39.0	44.6	26.8	0.5	3.0	19.6	41.7	17.0	42.3	3.4	**2.5**	0.0	15.2	24.5	12.7
CO-Occ [2]	59.7	35.2	21.2	1.7	39.3	44.1	28.3	0.5	3.1	21.4	41.5	18.1	42.5	3.4	2.3	0.0	14.5	24.2	12.2
MR-Occ	**62.3**	**39.7**	**26.4**	1.0	**40.4**	**46.6**	**31.3**	**4.2**	**6.8**	**21.5**	**43.8**	**26.3**	**47.1**	**6.5**	**2.5**	**0.0**	**19.0**	**28.4**	**17.3**

TABLE VI
SEMANTIC SCENE COMPLETION RESULTS ON SEMANTICKITTI VALIDATION SET. THE MODEL IS TRAINED ON SEMANTICKITTI TRAIN SET AND EVALUATED ON SEMANTICKITTI VALIDATION SET. ALL EXPERIMENTS WERE CONDUCTED USING LIDAR AND CAMERA INPUTS. THE BEST PERFORMANCE IN EACH CATEGORY IS HIGHLIGHTED IN BOLDFACE.

Methods	Metric	Standard	Sunny	Rain	Day	Night	CD_1	CD_3	CD_5	LBR_4	LBR_16
MCONet	IoU	29.5	33.5	27.2	34.3	22.3	24.0	18.8	12.6	24.8	22.0
	mIoU	20.1	23.3	21.5	23.9	10.8	18.4	13.5	6.8	20.0	19.0
MR-Occ	IoU	35.5	38.1	33.2	39.3	31.0	33.3	30.8	28.0	32.4	29.9
	mIoU	27.3	31.8	27.2	31.2	18.6	24.2	20.0	14.4	24.9	24.0

TABLE VII
ROBUSTNESS EVALUATION ACROSS VARIOUS CONDITIONS. CD_X INDICATES CAMERA DROP OF X VIEWS, AND LBR_Y DENOTES LIDAR BEAM REDUCTION TO Y BEAMS. RESULTS SHOW IoU AND mIoU METRICS (%).

	Background	Foreground
Coarse	93.4	6.6
Semi-fine	23.5	76.5
Fine	3.4	96.6

TABLE VIII
REGION DISTRIBUTION ANALYSIS OF RIE PREDICTIONS ACROSS DIFFERENT RESOLUTION LEVELS, DEMONSTRATING PROGRESSIVE FOCUS ON FOREGROUND REGIONS.

τ_1	τ_2	IoU	mIoU
0.7	0.4	35.1	26.5
0.5	0.7	35.3	27.0
0.4	0.8	35.1	27.0
0.4	0.6	35.3	27.2
0.4	0.7	**35.5**	**27.3**
0.3	0.7	35.4	27.1

TABLE IX
ABLATION STUDY ON THE THRESHOLDS FOR HVFE MODULE USING THE NUSCENES-OCCUPANCY VALIDATION SET. τ_1 AND τ_2 ARE THE THRESHOLDS FOR SELECTING THE SEMI-FINE RESOLUTION SET $\mathcal{S}$ AND THE FINE RESOLUTION SET $\mathcal{F}$.
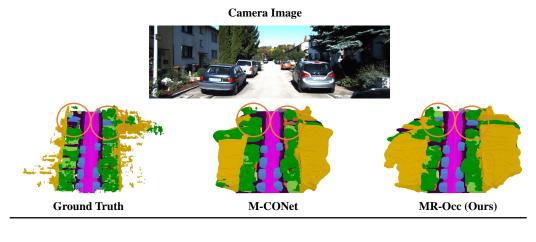
25.9% mIoU, demonstrating that densification plays a critical role in enhancing object boundary details and overall prediction accuracy.

Efficiency and Accuracy Analysis. Table IV compares the efficiency and accuracy of various models. Our proposed MR-Occ achieves state-of-the-art performance while requiring the least computational resources. Additionally, MR-Occ attains an inference speed of 3.2 FPS, the second fastest after M-OpenOccupancy. Note that our model significantly outperforms M-OpenOccupancy in accuracy, with a 6.4% higher IoU and a 12.2% higher mIoU, while maintaining competitive inference speed. The FPS difference stems from architectural trade-offs: while M-OpenOccupancy uses a computationally heavy Encoder but simple Occupancy Head, MR-Occ employs an efficient low-resolution Encoder but adopts M-CONet's Occupancy Head, which introduces latency through memory-intensive voxel refinement operations.

Robustness Analysis. We conduct extensive experiments to evaluate MR-Occ's robustness under various challenging conditions and sensor configurations. As shown in Table VII, we analyze performance across different weather conditions (Sunny, Rain, Day, Night) and sensor degradation scenarios (camera drop and LiDAR beam reduction). MR-Occ consis-

tently outperforms M-CONet across all test conditions. In weather scenarios, our model maintains strong performance with 38.1% IoU in sunny conditions and demonstrates resilience in challenging scenarios like rain (33.2% IoU) and night (31.0% IoU). Notably, MR-Occ shows robust adaptation to sensor limitations: even with significant camera view drops (CD_5) and LiDAR beam reductions (LBR_16), it achieves 28.0% IoU and 29.9% IoU respectively, outperforming M-CONet by substantial margins (15.4% and 7.9% respectively). These results demonstrate that our model's hierarchical feature refinement strategy effectively maintains performance even under degraded sensor conditions.

Resolution Importance Estimator Analysis. We demonstrate RIE's importance through a quantitative analysis by comparing the ratio of foreground and background regions in RIE's prediction results (Coarse voxel set (C), Semi-fine (S), and Fine Resolution (F)). As shown in the Table VIII, the majority of C corresponds to background regions (93.4%), while S and F increasingly correspond to foreground regions (76.5% and 96.6% respectively). This indicates that the model effectively performs hierarchical predictions through RIE, enabling it to
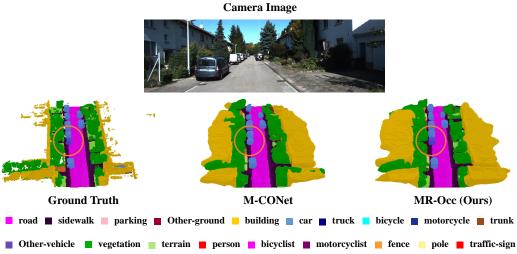
Camera Image

| Ground Truth | M-CONet | MR-Occ (Ours) |

Camera Image

| Ground Truth | M-CONet | MR-Occ (Ours) |

■ road ■ sidewalk ■ parking ■ Other-ground ■ building ■ car ■ truck ■ bicycle ■ motorcycle ■ trunk

■ Other-vehicle ■ vegetation ■ terrain ■ person ■ bicyclist ■ motorcyclist ■ fence ■ pole ■ traffic-sign

Fig. 5. Qualitative comparison results on SemanticKITTI validation set. The regions highlighted by orange circles indicate areas with obvious differences.

focus on foreground voxels. The progressive refinement strategy of RIE proves particularly effective, as it allows the model to allocate computational resources primarily to regions of interest. This hierarchical approach not only improves accuracy but also demonstrates computational efficiency by avoiding unnecessary processing of background regions at higher resolutions.

Hyperparameter Analysis. Distinguishing key voxels in the Semi-fine Resolution set S and Fine Resolution set F is crucial for enhancing feature representation in the Hierarchical Voxel Feature Refinement (HVFR) module. We conducted an extensive ablation study to identify the optimal thresholds τ_1 and τ_2 for voxel selection, as summarized in Table IX.

Different combinations of τ_1 and τ_2 were evaluated to assess their impact on Intersection over Union (IoU) and mean Intersection over Union (mIoU) metrics. The results demonstrate that these thresholds significantly influence the model's performance in voxel feature refinement. Among the tested configurations, setting $\tau_1 = 0.4$ and $\tau_2 = 0.7$ achieves the highest IoU of 35.5 and mIoU of 27.3. This threshold combination strikes an optimal balance by effectively capturing key voxels across different resolution scales, thus improving overall feature enhancement.

It was observed that lower values of τ_1 (e.g., 0.3) or higher values of τ_2 (e.g., 0.8) lead to marginally reduced performance. This reduction is likely due to the over- or under-selection of key voxels, which affects the balance between the Semi-fine and Fine Resolution sets. Despite these variations, the IoU and mIoU values remain relatively consistent across different threshold combinations, indicating the robustness of the HVFR module to small changes in threshold values.

Quantitative Results on the SemanticKITTI test set. We present a comparative analysis of MR-Occ's performance against reproduced results from existing models on the SemanticKITTI test set in Table V. MR-Occ shows an improvement of 0.8% in IoU and 2.7% in mIoU compared to the baseline model, M-CONet. When compared to Co-Occ, MR-Occ achieves a 0.9% higher IoU and a 3.2% higher mIoU. Our proposed model consistently demonstrates outstanding performance across various datasets.

Table VI provides a comprehensive evaluation of semantic scene completion performance on the SemanticKITTI validation set. MR-Occ consistently outperforms existing methods across various semantic classes, underscoring its effectiveness in complex urban environments. In critical classes for autonomous driving applications, MR-Occ demonstrates significant gains. For instance, in the 'car' class, our model achieves 46.6%

accuracy, surpassing M-CONet by 2.0% and CO-Occ by 2.5%. Similarly, MR-Occ records a 4.5% improvement over M-CONet for the 'truck' class. In challenging classes like 'motorcycle' and 'person,' our model outperforms M-CONet by 3.8% and 3.1%, respectively, demonstrating its superior capability in handling complex detection tasks. These performance gains underline MR-Occ's capability to surpass existing SOTA methods, achieving more accurate and comprehensive semantic scene completion.

E. Qualitative Results

nuScenes-Occupancy. Figure 4 provides a visual comparison of 3D semantic occupancy predictions between our proposed MR-Occ and the baseline M-CONet on the nuScenes-Occupancy dataset. MR-Occ exhibits superior performance in capturing fine-grained details and accurately predicting occluded regions across various urban scenarios. In the first scene, MR-Occ accurately delineates the truck and vegetation, preserving sidewalk continuity. The second scene showcases our model's precision in segmenting multiple cars in close proximity, maintaining clear boundaries between them. In the third scene, MR-Occ excels in predicting complex structures like the bus stop, accurately capturing glass panels and pedestrians often missed by M-CONet. These results highlight MR-Occ's effectiveness in producing more accurate and detailed 3D semantic occupancy predictions, particularly in challenging urban environments with multiple object classes and occlusions.

SemanticKITTI. Figure 5 presents a visual comparison of 3D semantic occupancy predictions between MR-Occ and the M-CONet baseline on the SemanticKITTI validation set. The results clearly highlight MR-Occ's superior performance, particularly in complex urban environments.

MR-Occ effectively delineates the boundary between sidewalks and adjacent regions, providing precise predictions. It also excels in predicting the position and shape of distant and partially occluded cars, delivering reliable results even where M-CONet struggles. Additionally, MR-Occ accurately identifies object classes in complex environments where various objects are intermingled. These results demonstrate that MR-Occ effectively leverages multimodal data, ensuring robust performance across various urban environments.

V. CONCLUSIONS

We introduce MR-Occ, a novel and efficient camera-LiDAR fusion method for 3D semantic occupancy prediction. MR-Occ excels by utilizing a Pixel to Voxel Fusion Network, Hierarchical Voxel Feature Refinement, and a Multi-scale Occupancy Decoder, effectively addressing key challenges such as sensor misalignment and the accurate prediction of occluded regions. Our method achieves state-of-the-art performance on the nuScenes-Occupancy dataset and demonstrates highly competitive results on the SemanticKITTI dataset. This is accomplished with fewer parameters and reduced computational complexity, establishing MR-Occ as a highly efficient solution. Future work will focus on integrating temporal information to further enhance stability in dynamic environments. We hope that MR-Occ serves as a strong baseline in camera-LiDAR

fusion for 3D semantic occupancy prediction, contributing valuable insights to future research in this area.

REFERENCES

[1] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 850–17 859.

[2] J. Pan, Z. Wang, and L. Wang, "Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction," *IEEE Robotics and Automation Letters*, 2024.

[3] G. Wang, Z. Wang, P. Tang, J. Zheng, X. Ren, B. Feng, and C. Ma, "Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving," *arXiv preprint arXiv:2404.15014*, 2024.

[4] L. Roldao, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 111–119.

[5] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3cnet: A sparse semantic scene completion network for lidar point clouds," in *Conference on Robot Learning*. PMLR, 2021, pp. 2148–2161.

[6] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109.

[7] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.

[8] Z. Xia, Y. Liu, X. Li, X. Zhu, Y. Ma, Y. Li, Y. Hou, and Y. Qiao, "Scpnet: Semantic scene completion on point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 642–17 651.

[9] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2308.16896*, 2023.

[10] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3991–4001.

[11] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9433–9443.

[12] Z. Yu, C. Shu, J. Deng, K. Lu, Z. Liu, J. Yu, D. Yang, H. Li, and Y. Chen, "Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin," *arXiv preprint arXiv:2311.12058*, 2023.

[13] J. Hou, X. Li, W. Guan, G. Zhang, D. Feng, Y. Du, X. Xue, and J. Pu, "Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view," *arXiv preprint arXiv:2403.02710*, 2024.

[14] P. Tang, Z. Wang, G. Wang, J. Zheng, X. Ren, B. Feng, and C. Ma, "Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 035–15 044.

[15] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9223–9232.

[16] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9087–9098.

[17] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[18] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[19] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[20] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[21] Y. Man, L.-Y. Gui, and Y.-X. Wang, "Bev-guided multi-modality fusion for driving perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 960–21 969.

[22] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4413–4421.

[23] A. Geiger and U. R. Lenzp, "Arewereadyfor autonomousdriving," 2012.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[26] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[27] M. Contributors, "Mmdetection3d: Openmmlab next-generation platform for general 3d object detection," 2020.

[28] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[29] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

BIOGRAPHY SECTION

Hawook Jeong received his Ph.D. degree in Electrical and Computer Engineering from Seoul National University, Seoul, South Korea, in February 2015. He also received his M.S. degree in Electrical and Computer Engineering in 2011 and his B.S. degree in Electrical Engineering in 2009, both from Seoul National University. Since October 2018, he has been with RideFlux, where he has served as a research scientist overseeing the development of perception and artificial intelligence algorithms for autonomous driving.

Jun Won Choi earned his B.S. and M.S. degrees from Seoul National University and his Ph.D. from the University of Illinois at Urbana-Champaign. Following his studies, he joined Qualcomm in San Diego, USA, in 2010. From 2013 to 2024, he served as a faculty member in the Department of Electrical Engineering at Hanyang University. Since 2024, he has held a faculty position in the Department of Electrical and Computer Engineering at Seoul National University. He currently serves as an Associate Editor for both IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Vehicular Technology, International Journal of Automotive Technology. His research spans diverse areas including signal processing, machine learning, robot perception, autonomous driving, and intelligent vehicles.

Minjae Seong received the B.S. degree in Automotive Engineering in 2020 and the M.S. degree in Artificial Intelligence in 2023, both from Hanyang University, Seoul, South Korea. He is currently pursuing the Ph.D. degree in Artificial Intelligence at Hanyang University. His research interests include deep learning-based multi-modal 3D perception, computer vision, robot perception, and autonomous driving.

Jisong Kim received the B.S. degree in Automotive Engineering in 2020 and the M.S. degree in Electrical Engineering in 2022, both from Hanyang University, Seoul, South Korea. He is currently pursuing the Ph.D. degree in Electrical Engineering at Hanyang University. His research interests include multi-modal 3D perception, deep learning, sensor fusion, knowledge distillation, and autonomous driving.

Geonho Bang received the B.S. degree in Automobile and IT Convergence in 2022 from Kookmin University, Seoul, South Korea and the M.S. degree in Artificial Intelligence from Hanyang University, Seoul, South Korea, in 2025. He is currently pursuing the Ph.D. degree in the Interdisciplinary Program in Artificial Intelligence at Seoul National University, Seoul, South Korea. His research interests include multi-modal 3D perception, sensor fusion, knowledge distillation, and autonomous driving.