2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

TACOcc: Target-Adaptive Cross-Modal Fusion with Sequential Volume Rendering for 3D Semantic Occupancy Prediction

Luyao Lei, Shuo Xu, Yifan Bai†, Zelin Yang, Yuanbo Guo, and Xing Wei*

Abstract— **Multi-modal 3D semantic occupancy prediction remains challenged by two fundamental issues: (i) geometric–semantic misalignment introduced by fixed-neighborhood fusion under heterogeneous sensing distributions, and (ii) feature degradation with prediction inconsistency in dynamic scenes caused by sparse supervision. We propose *TACOcc*, a framework coupling a target-adaptive, bidirectional symmetric fusion module with sequential volume rendering supervision. The fusion module predicts a query-wise neighborhood size via a differentiable Gumbel-Softmax strategy, expanding the receptive field for large objects to enrich context while contracting it for small objects to suppress noise, thereby achieving precise cross-modal alignment. To stabilize predictions under sparse labels and motion, we introduce temporally enhanced Gaussian rendering that aggregates multi-frame dependencies, initializes dual-source geometric anchors, and transfers multi-view photometric constraints from images to 3D occupancy features. A velocity-adaptive temporal bandwidth further mitigates flicker in fast-motion cases. Experiments on nuScenes and SemanticKITTI demonstrate strong performance, including 28.9% mIoU on nuScenes, particularly improving small-object categories and long-range regions. These results highlight that scale-aware bidirectional fusion and temporally grounded volumetric supervision form an effective recipe for robust multi-modal occupancy perception.**

I. INTRODUCTION

3D semantic occupancy prediction provides fine-grained geometric and semantic representations for real-time environmental perception in autonomous driving by densely parsing voxelized 3D space [1]–[8]. Compared to traditional 3D object detection constrained by rigid bounding box representations, this technique effectively models irregular obstacles (e.g., fallen vegetation or construction barriers) and continuous surface structures (e.g., road undulations). It demonstrates unique advantages in critical tasks such as dynamic obstacle avoidance and drivable area segmentation [8]–[10]. However, current methods face two major bottlenecks in leveraging multi-modal data for enhanced scene understanding: (1) Cross-modal fusion mechanisms based on fixed sampling neighborhoods cause spatial misalignment between geometric features and semantic representations [11]–[14]; (2) Sparse annotations lead to feature degradation and prediction inconsistency in dynamic scenes [15].

To address geometric-semantic misalignment, recent studies employ attention mechanisms for dynamic modality weighting [16]. Yet these methods perform global filtering only along feature channels, failing to capture spatially varying local correlations. For feature degradation, emerging 3D Gaussian Splatting (3DGS) [17] leverages photometric constraints to optimize features and improve prediction accuracy, but its uni-modal architecture limits cross-modal feature complementarity. Crucially, most existing methods ignore the temporal consistency of dynamic objects—a fundamental requirement in real-world driving scenarios.

To overcome these limitations, we propose ***TACOcc***, a novel 3D semantic occupancy prediction framework. For geometric-semantic misalignment, we design an *adaptive feature fuser* that dynamically predicts optimal neighborhood ranges for each query feature via a Gumbel-Softmax-optimized selection mechanism: expanding receptive fields for large objects to capture global context while contracting neighborhoods for small objects to suppress noise. This adaptive control guides a bidirectional cross-modal interaction unit to achieve precise multi-modal feature alignment. For feature degradation and inconsistency in dynamic scenes, we introduce a *sequential volume renderer*, which first employs a temporal consistency aggregator to establish spatio-temporal dependencies along the time dimension. It subsequently performs multi-modal temporal-enhanced Gaussian rendering and leverages input images to construct a learning pathway from 2D supervision signals to 3D feature space. This approach mitigates the scarcity of sparse annotations by exploiting photometric consistency supervision. Through explicit temporal dependency modeling and multi-frame information integration, it enhances prediction consistency and coherence. Furthermore, the prediction accuracy is improved through implicit multiple corrections across different frames.

We conduct comprehensive experiments on the nuScenes and SemanticKITTI benchmarks to demonstrate the effectiveness of the proposed method. The TACOcc method achieves a 28.9% mIoU in 3D semantic occupancy prediction, surpassing the current SOTA multi-modal method, CoOcc, by 2.3%. This confirms the effectiveness of the dual-module collaborative framework and provides a scalable new framework for multi-modal autonomous driving perception.

Our contributions are as follows: (1) We design a target-adaptive bidirectional symmetric retrieval that selects the optimal neighborhood per query via a differentiable Gumbel-Softmax/STE pathway, achieving precise alignment between geometric and semantic features across modalities and scales. (2) We introduce a sequential Gaussian renderer coupling temporal aggregation with dual-source geometric anchors, transferring multi-view photometric constraints into 3D oc-

*Corresponding author: weixing@mail.xjtu.edu.cn.

†Project leader. All authors are with Xi'an Jiaotong University, Xi'an, 710049, China. This work was supported by the National Natural Science Foundation of China No. 62572385, the Fundamental Research Funds for the Central Universities No. xxj032023020, and CAAI-CANN Open Fund, developed on OpenI Community.

cupancy space to compensate for sparse annotations. (3) A velocity-adaptive temporal bandwidth and parameter-consistency regularization mitigate flicker and stabilize predictions in dynamic scenes. TACOcc achieves state-of-the-art performance on nuScenes and strong results on SemanticKITTI, with notable gains for small objects and long-range perception.

II. RELATED WORK

3D Semantic Occupancy Prediction. The core challenge lies in effectively representing spatial occupancy states and their semantic categories in 3D scenes. Early works focused on indoor scenes [18], [19], later shifting to outdoor environments using sparse lidar data [20], [21]. MonoScene [1] pioneered monocular image-based occupancy prediction. Recent methods leverage multi-modal fusion for improved performance [22], [23]. However, static fusion mechanisms struggle with generalization. Co-Occ [22] uses a fixed neighborhood search, failing to align geometric and semantic features effectively. To address this, we introduce a target-adaptive bidirectional retrieval mechanism via joint optimization of Straight-Through Estimator(STE) and Gumbel-Softmax, enabling precise cross-modal matching.

Volume Rendering for Scene Understanding. Volume rendering improves geometric fidelity through photometric supervision of volumetric representations. Traditional methods lack cross-scene adaptability [24], [25]. Although [26], [27] investigated rendering-based occupancy prediction, they did not incorporate semantic information. Neural Radiance Fields (NeRF) [28] have been used for semantic occupancy prediction, which improves anti-aliasing and training efficiency [29], [30]. However, rendering speed and dynamic inconsistency remain major challenges. Recent point-based methods such as 3D Gaussian Splatting (3DGS) [17] achieve real-time rendering, but still face issues like uni-modal dependency, heuristic initialization, and temporal instability. Our approach introduces a multi-modal time-aware Gaussian renderer with automatic initialization and consistency mechanisms, effectively solving feature degradation and inconsistency problems.

III. METHOD

We use lidar point clouds and their surround-view images as inputs to predict 3D semantic occupancy in driving scenarios. Our framework consists of two key components: the adaptive feature fuser and the sequential volume renderer. As shown in Fig.1:

This framework first generates sky region masks using an unsupervised semantic segmentation algorithm to suppress non-informative areas in images, while applying conditional filtering to denoise raw lidar point clouds. The denoised point cloud data is voxelized, and a sparse convolutional neural network extracts voxel features. Simultaneously, the masked images are fed into an image encoder for feature extraction, with camera parameters and depth information enabling 2D-3D view transformation to project image features into the same 3D space as the point clouds. Then, we use the scale-adaptive fuser that predicts adaptive neighborhood ranges based on non-zero query features. These ranges guide a symmetric retrieval unit to enable bidirectional cross-modal interaction. The module dynamically adjusts aggregation weights between image and point cloud features, achieving complementary enhancement of geometric and semantic information. We then inject the fused features into the sequential volume renderer, establishing an optimization pathway from 2D supervision signals to the 3D feature space. Finally, the enhanced fused features are input into a cascaded occupancy head [23] to output a fine-grained 3D semantic occupancy grid.

A. Adaptive Feature Fuser

The adaptive fuser achieves precise multi-modal feature alignment in complex driving scenarios through target-adaptive bidirectional symmetric retrieval (illustrated in Fig 2), the alignment is achieved because the adaptive neighborhood selection enables the retrieval range, generated through features from the other modality, to match the object scale, avoiding noise or missing context introduced by a fixed neighborhood, while the bidirectional retrieval forces the geometric and semantic features to guide and constrain each other spatially, thereby establishing precise cross-modal correspondence.

Adaptive Neighborhood Selector. To enhance hierarchical and comprehensive cross-modal feature fusion, we first construct multi-scale pyramid features for the image features F_I and point cloud features F_L. A multilayer perceptron (MLP) then maps the current modality query features to a candidate integer k, which denotes the number of neighbors in the K-nearest neighbors (KNN) search within the bidirectional symmetric retrieval mechanism. The optimal k may vary across different scales. The core challenge in optimal k selection lies in parameterized modeling of discrete probability distributions: The non-differentiability of traditional discrete sampling operators (e.g., argmax) blocks gradient backpropagation. To address this, we introduce the Gumbel-Softmax reparameterization technique [31] to establish a continuous relaxation mechanism. The generated probability distribution is multiplied by candidate k values to compute the expectation, followed by a rounding operation to determine the optimal k for the current scene.

However, the rounding discretization operation disrupts computational graph integrity, causing gradient backpropagation chain breakage at critical decision nodes. We employ the Straight-Through Estimator (STE) [32] to construct a proxy gradient function, decoupling forward/backward propagation gradient computation. The combination of STE and Gumbel-Softmax effectively resolves the discretization issue of argmax-based optimal k selection while preserving gradient propagation.

Multi-Scale Pyramid Fuser (MSPF). Utilizes attention mechanisms to extract key information from multi-scale pyramid features. At each scale, KNN search is performed for every non-zero query feature in the current modality (image or point cloud), retrieving the top-k non-zero voxel

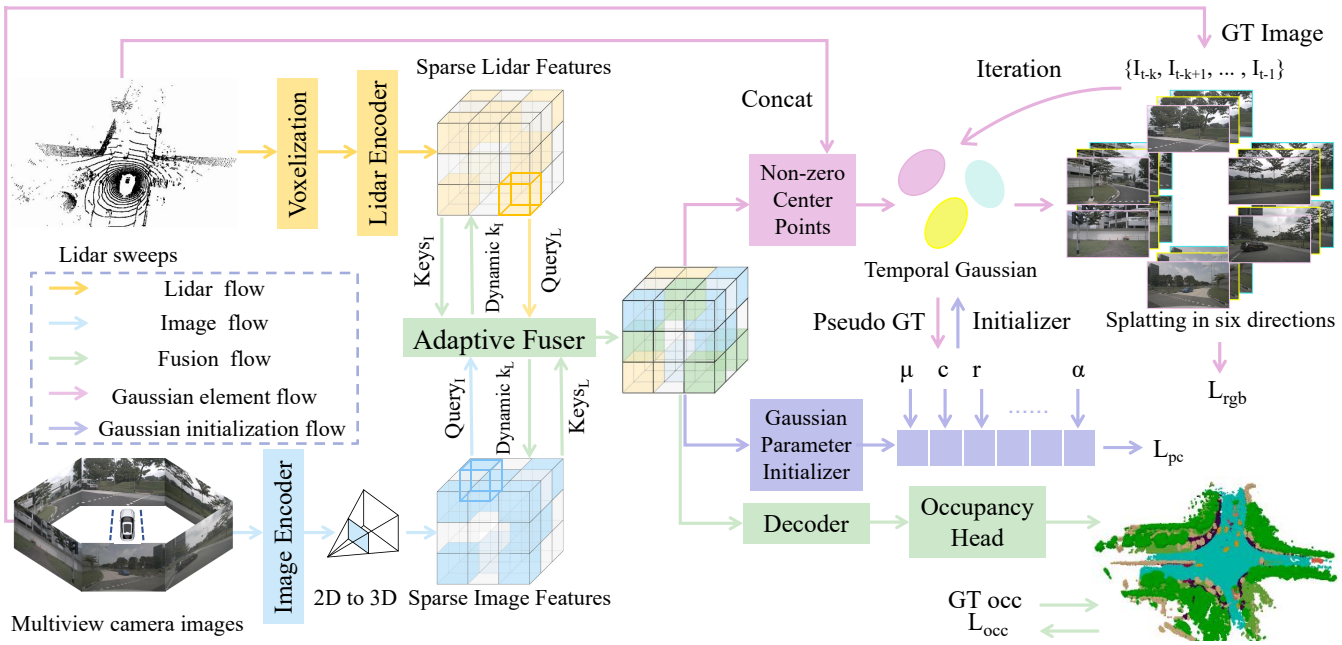

Fig. 1. **Overview of proposed TACOcc.** The point clouds and images undergo feature extraction and dimensional transformation to generate sparse voxel features. These features are then precisely aligned and fused through an adaptive feature fuser (See Section III-A), followed by enhancement via a sequential volume renderer (See Section III-B), thereby improving the performance of 3D semantic occupancy prediction.

key vectors. These vectors are concatenated and fed into an attention module, enabling the model to autonomously learn weights for the current query. The weights for all queries across different modalities are denoted as ω_I or ω_L. The weights are multiplied with corresponding query features to achieve comprehensive aggregation of multi-scale contextual information (Formula 1), addressing the lack of hierarchical semantics in single-scale feature fusion.

$$F_{IL} = \text{Concat}(F_I, F_L, F_I \cdot \omega_I, F_L \cdot \omega_L) \tag{1}$$

B. Sequential Volume Renderer

We employ sequential volume rendering techniques (with the sequence window size T set to 5) to address feature degradation and prediction inconsistency in dynamic scenes caused by sparse annotations, while further refining the fused features.

Temporal Consistency Aggregator. To resolve prediction inconsistency, a temporal consistency aggregator is designed using consecutive frame data. The fused features $F_{IL} \in \mathbb{R}^{T\times H\times W\times D\times C}$ are processed by a 3D ConvLSTM layer:

$$\tilde{F}^t_{IL} = \text{ConvLSTM}(F^t_{IL}, \tilde{F}^{t-1}_{IL}) \tag{2}$$

It establishes spatiotemporal feature dependencies. Multiple corrections across frames significantly enhance rendering accuracy while ensuring stability.

Cross-Modal Geometric Anchor Coupling. To address point cloud sparsity, we couple dual-source geometric anchors (DGA) as Gaussian centers: original point cloud coordinates and centroids of non-zero voxels from temporally smoothed features $\tilde{F}^t_{IL}$. Point clouds provide precise geometry and constrain offsets, while voxel grids compensate for sparsity, eliminating uni-modal blind spots.

Multi-modal Temporally Enhanced Gaussian. To address temporal consistency and rendering smoothness in 3D Gaussian modeling for dynamic scenes, each Gaussian primitive incorporates a temporal dimension t and temporal weights $\mathcal{N}(t|\mu_t, \sigma_t^2)$, enabling explicit temporal parameterization and smooth inter-frame interpolation. The extended parameters are (Formula 3):

$$\mathcal{G}_i = \{\mu_i, \mathbf{r}_i, \mathbf{s}_i, \mathbf{c}_i, \alpha_i, \mu_t, \sigma_t\} \tag{3}$$

where μ_i is the 3D position center, $\mathbf{r}_i$ is the rotation quaternion, $\mathbf{s}_i$ is the scale factor, $\mathbf{c}_i$ are spherical harmonic coefficients, α_i is opacity, μ_t is the temporal center, and σ_t is the temporal bandwidth.

During parameter initialization, a lightweight convolutional network processes the temporally smoothed features $\tilde{F}_{IL}$ to obtain initial parameters for each Gaussian component. Differentiable rendering techniques [17] project 3D Gaussians onto 2D image space across six views, with temporal weights enabling natural inter-frame motion interpolation. Pixel color computation adopts the neural point-based method [37], blending ordered points overlapping each pixel. To suppress flickering artifacts caused by rapid motion, a velocity-adaptive Gaussian filter is designed: Activated when voxel velocity $\|\mathbf{v}\|$ (displacement vector computed via adjacent frame point cloud registration) exceeds a scene-dependent threshold τ_{velocity} (Formula 4):

$$\tau_{\text{velocity}} = \frac{\beta}{\Delta t} \cdot \mathbb{E}[\|\mathbf{v}_{\text{scene}}\|] \tag{4}$$

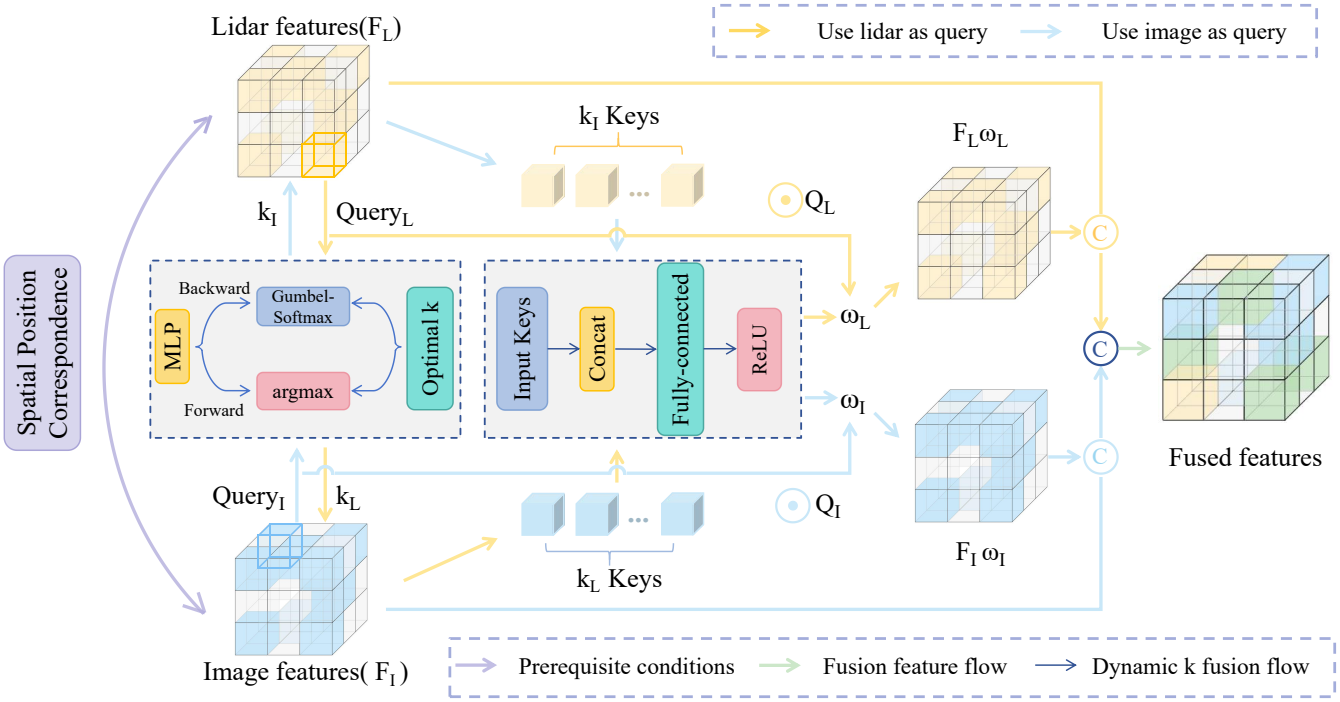

Fig. 2. **Adaptive Feature Fuser.** For target-aligned fusion of bimodal sparse voxel features, the adaptive neighborhood selector generates optimal k values for each query at different scales. Bidirectional symmetric retrieval is executed based on these values. Retrieved key vectors are stacked and processed through feature extraction and nonlinear transformation to compute attention weights, which ultimately weight the current query features to produce fused features.

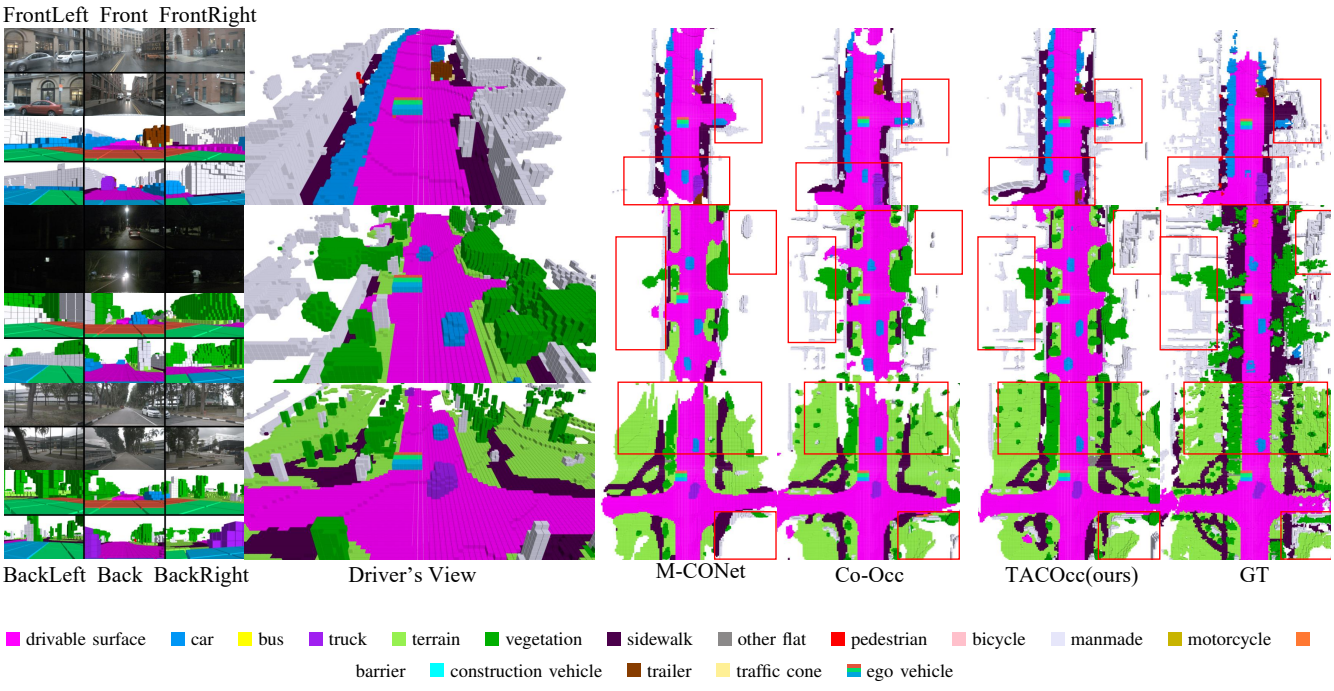

Fig. 3. Qualitative comparison results on the nuScenes validation set. (Left) Input images and TACOcc's per-image predictions. (Right) Driver view (Ours) and BEV comparisons: M-CONet [23], Co-Occ [22], Ours, and GT [8].

where β is a learnable scaling factor and Δt is the frame interval. Upon activation, the temporal bandwidth σ_t of moving Gaussians is temporarily increased to smooth their inter-frame appearance changes, suppressing flickering artifacts from rapid motion while preserving static background details.

TABLE I

3D SEMANTIC OCCUPANCY PREDICTION RESULTS ON NUSCENES VALIDATION.

Method	Modality	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation	Input Size	2D Backbone
MonoScene [1]	C	22.4	7.2	6.4	2.2	8.6	7.1	4.5	2.3	2.8	1.9	4.0	4.4	16.3	6.2	12.5	10.3	9.7	15.8	900×1600	R101-DCN
SurroundOcc [8]	C	31.5	20.2	19.3	11.8	29.1	10.0	15.4	13.9	12.3	14.1	21.3	37.3	24.7	24.2	17.4	23.3	23.9	24.9	900×1600	R101-DCN
BEVFormer [33]	C	30.7	16.8	13.9	6.3	23.8	27.9	8.3	10.4	6.9	4.8	10.8	18.2	38.1	19.2	22.7	21.6	14.1	22.0	900×1600	R101-DCN
C-CONet [23]	C	25.7	18.6	18.7	10.3	27.8	27.7	8.2	16.3	13.8	9.6	11.3	19.2	33.2	20.5	22.1	21.6	14.9	22.6	896×1600	R101
OccFormer [34]	C	30.1	20.5	29.8	11.5	28.4	30.8	10.3	16.3	13.9	12.6	14.6	20.7	36.8	22.3	23.5	22.1	14.2	20.6	896×1600	R101
RenderOcc [29]	C	29.0	18.7	18.9	11.3	27.9	9.9	13.7	13.8	11.9	12.7	20.5	32.5	21.2	24.0	21.5	22.1	15.2	21.6	896×1600	R101
LMSCNet [35]	L	32.6	13.2	12.8	4.2	13.4	19.1	10.6	4.9	7.3	7.5	10.4	11.3	23.3	12.7	16.5	13.9	15.1	28.0	–	–
L-CoNet [23]	L	39.1	18.1	19.0	3.9	16.2	27.3	7.3	4.1	7.2	5.8	14.6	14.3	38.9	20.2	24.5	24.2	25.1	36.9	–	–
M-CoNet [23]	C&L	39.2	23.6	23.4	12.4	31.1	33.2	14.1	17.2	18.7	13.4	19.9	25.2	38.2	21.2	25.4	25.1	24.7	34.9	896×1600	R101
LC-Fusion [36]	C&L	40.7	25.0	27.8	16.3	**33.7**	33.2	17.5	17.2	18.7	15.4	19.5	25.2	39.1	21.2	22.3	**29.1**	**29.6**	34.9	896×1600	R101
Co-Occ [22]	C&L	41.0	26.6	27.3	16.5	33.5	37.5	17.9	21.8	16.8	21.7	**27.6**	39.2	25.9	**28.1**	29.2	26.2	18.9	36.8	896×1600	R101
TACOcc (Ours)	C&L	**44.2**	**28.9**	**30.1**	**17.5**	33.2	**38.3**	**18.2**	**22.1**	**19.8**	**23.2**	25.4	**40.3**	**40.8**	27.8	**30.8**	28.7	29.1	**37.6**	896×1600	R101

Rendered images supervise Gaussian primitive optimization under input image sequences. A view-space position gradient-based adaptive density control strategy drives optimization, guiding adaptive growth along geometric gradients. Model parameters are iteratively optimized from initial values using multi-view photometric consistency loss $\mathcal{L}_{rgb}$, dynamically adjusting each Gaussian's radiative properties and geometry. Once converged, parameters serve as pseudo-GT and are compared with initial parameters via parameter consistency loss $\mathcal{L}_{pc}$, compensating for missing 2D details in fused 3D multi-modal features.

C. Optimization

Occupancy Loss. Use cross-entropy loss ($\mathcal{L}_{\text{ce}}$) and Lovász-softmax loss to supervise the relationship between predicted semantic occupancy and GT occupancy [22]: $\mathcal{L}_{\text{occ}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{lovasz}}$.

Gaussian Parameter Consistency Loss. Calculated from the initial and optimized Gaussian parameters [38]: $\mathcal{L}_{pc} = \sum_{i=1}^{N} \left\| \Theta_i^{\text{final}} - \Theta_i^{(0)} \right\|_1$.

Volume Rendering Photometric Loss. A combination of L1 and D-SSIM used to supervise rendered and real images [38]: $\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D\text{-}SSIM}$.

Total Loss. During Gaussian parameter iteration, $\mathcal{L}_{\text{rgb}}$ updates only Gaussians; post-iteration, weighted $\mathcal{L}_{\text{occ}}$ and $\mathcal{L}_{\text{pc}}$ refine fused features for enhanced detail.

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{rgb}} & \text{During Iteration} \\ (1-\lambda)\mathcal{L}_{\text{occ}} + \lambda\mathcal{L}_{\text{pc}} & \text{After Iteration} \end{cases} \tag{5}$$

IV. EXPERIMENTS

V. DATASET

NuScenes [39]. Supports voxelized occupancy prediction via multi-modal spatio-temporal alignment. Includes 1000 scenes (850 train, 150 val) with synchronized LiDAR, 6 cameras, and radar. Advantages over KITTI: dense 3D annotations for 23 classes, 20s trajectories with motion compensation, multi-modal alignment, and pose/IMU data. Ideal for dynamic urban occupancy.

SurroundOcc [8]. Extends nuScenes with automated dense occupancy labels via multi-frame fusion and Poisson voxelization (0.5m, 200×200×16). Offers full nuScenes compatibility, automated annotation, and superior density over Occ3D-nuScenes and SemanticKITTI. Suitable for fusion, small objects, and real-time benchmarks.

SemanticKITTI [40]. Provides dense point-wise annotations and sequential LiDAR data across 22 sequences. Features: 20-class labels, 360° coverage, accurate odometry, and challenging scenes. Key benchmark for geometric-detailed occupancy tasks.

A. Implementation Details

We evaluated the 3D semantic occupancy performance on the nuScenes validation set [39]. Using a ResNet101 [41] backbone and a 2D-to-3D view transformer [42] to generate a 3D feature volume of size $128\times128\times16$. The 10 lidar point clouds were voxelized using a voxel encoder. The adaptive feature fuser fused the features with candidate k values of 1, 2, 3. The same occupancy decoder and head as in [23] were used with a cascade ratio of 2 for refined predictions. The model was trained for 15 epochs on nuScenes using 8 RTX A6000 GPUs, completing training in 2 days. The AdamW [43] optimizer with a weight decay of 0.01 and an initial learning rate of 1×10^{-4}, using a stepwise cosine decay, was employed. The D-SSIM window size was 11×11, and λ was set to 0.2 in all experiments.

B. Experimental Results

We conduct experiments on the nuScenes [39] dataset and compared our approach with several SOTA methods across different modalities (Tab.I), including camera-only methods [1], [8], [23], [29], [33], [34], lidar-only methods [23], [35], and lidar-camera fusion methods [22], [23]. Our results show that the proposed TACOcc method achieves a 3D semantic occupancy prediction mIoU of 28.9%, outperforming the previous best multi-modal method, Co-Occ [22],

by 2.3%. It improves geometric-semantic consistency for small objects such as pedestrians and cyclists. Specifically, it improves the IoU of sparse point cloud objects such as traffic cones and motorcycles by nearly 3%, while maintaining a high IoU for medium and large vehicles.

We present visualization results on the nuScenes dataset in Fig.3, demonstrating our method's significant advantages in both completeness and detail of 3D occupancy prediction.

To further validate the effectiveness of our framework, we conduct a comparative analysis with SOTA methods on the SemanticKITTI test set [40] (Tab.II). As shown in the table, our method outperforms JS3CNet [21] by 3.5% mIoU and SSC-RS [44] by 2.1% mIoU, despite their use of additional lidar segmentation supervision.

C. Ablation Studies

Through ablation studies, the TACOcc framework systematically validates the synergistic effect of dynamic multi-modal fuser and sequential volume rendering supervision for 3D semantic occupancy prediction. Experimental results (Table III) show that the baseline model achieves 34.3% IoU and 22.1% mIoU on the *nuScenes* dataset. Using fixed-neighborhood sampling fuser ($k = 3$) causes target misalignment and introduces noise. After incorporating target-scale adaptive multiscale retrieval, performance improves to 40.9% IoU (+6.6%) and 25.6% mIoU (+3.5%), confirming improved alignment. Furthermore, a differentiable transformation pipeline maps 3D features to multi-view 2D spaces. By applying both the photometric reprojection loss $\mathcal{L}_{\text{rgb}}$ and the Gaussian consistency loss $\mathcal{L}_{\text{pc}}$, a bidirectional 2D-3D supervision scheme is formed. The model finally reaches 44.2% IoU and 28.9% mIoU, demonstrating that dual supervision enhances both perception and reconstruction.

We evaluate the impact of the neighborhood retrieval range (Tab.IV). Experimental results show that the fixed k strategy improves accuracy but reduces efficiency rapidly. However, the dynamic range strategy achieves a balance between accuracy and efficiency. When k is set to 1-3, the IoU and mIoU reach 44.2% and 28.9%, respectively. This represents an improvement of 3.0% and 2.8% compared to fixed k = 3, while the latency remains low at only 0.6s on a single RTX A6000 GPU.

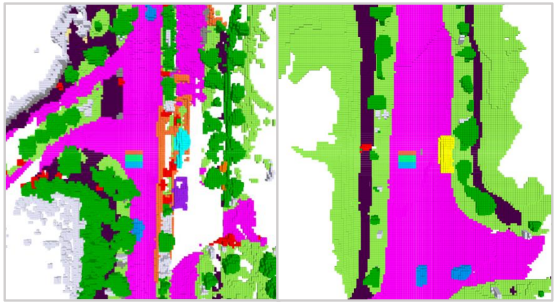

Fig. 4. Scenarios with more small targets compared with those dominated by large targets.

A comparison of the distribution of k in different target scenarios shows that in scenarios (Tab. V, Fig. 4) dominated by small targets, the proportion of k = 1 reaches 30.2%, indicating that small targets tend to use a narrow neighborhood to suppress noise. In contrast, in scenarios with large targets, the proportion of k = 3 accounts for 39.4% while that of k = 1 is only 16.9%, suggesting that large targets require an expanded neighborhood to enhance context understanding. This difference validates the effectiveness of the scale-perception mechanism in the dynamic strategy.

We compares the performance of different methods in 3D semantic occupancy prediction at distances from 25m to 100m (Tab.VI). TACOcc outperforms other methods in both short and long-range scenarios, demonstrating strong performance. The results confirm the effectiveness of its dynamic approach in complex environments.

We evaluated the inference time and memory usage of different image sizes and 2D backbone networks on a single RTX A6000 GPU (Tab.VII). For the model with low image resolution and ResNet50, its computational cost and latency are low, but its performance is not high. Increasing the image size and deepening the ResNet do not significantly increase the memory usage and latency, and can greatly improve the performance.

We compare rendering results from the sequential volume renderer in Fig. 5, where real input images (top) and volume-rendered outputs (bottom) are shown. The rendered results exhibit high fidelity in geometric detail, surface reflectance, and lighting consistency.

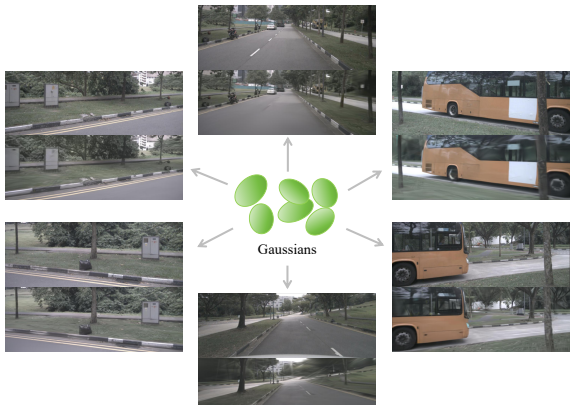

Fig. 5. Rendered images display and comparison with real images. For each viewpoint, the image above is the real one, and the image below is the rendered result.

VI. CONCLUSIONS

We presented TACOcc, a multi-modal occupancy framework that jointly tackles cross-modal misalignment and temporal inconsistency. Our target-adaptive bidirectional symmetric fusion delivers scale-aware alignment, while sequential (temporally enhanced) Gaussian rendering transfers rich photometric signals into 3D, compensating for label sparsity and stabilizing dynamics. TACOcc achieves state-of-the-art results on nuScenes and strong performance on

TABLE II

3D SEMANTIC OCCUPANCY PREDICTION RESULTS ON SEMATICKITTI TEST SET.

Method	Modality	mIoU	**road** (13.90%)	**sidewalk** (11.3%)	**parking** (1.12%)	**other ground** (0.56%)	**building** (14.1%)	**car** (9.92%)	**truck** (0.64%)	**bicycle** (0.03%)	**motorcycle** (0.03%)	**other vehicle** (0.20%)	**vegetation** (39.3%)	**trunk** (0.51%)	**terrain** (1.79%)	**person** (0.07%)	**bicyclist** (0.07%)	**motorcyclist** (0.03%)	**fence** (3.09%)	**pole** (0.29%)	**traffic sign** (0.08%)
MonoScene [1]	C	10.9	54.2	26.4	24.1	4.8	14.2	18.7	3.6	0.7	0.9	4.2	15.3	2.7	19.2	1.6	0.2	0.3	11.5	3.1	2.0
SurroundOcc [8]	C	11.7	56.1	27.9	31.1	6.3	14.8	21.0	1.2	1.3	1.1	4.6	15.1	4.3	19.1	1.5	2.2	0.2	10.1	4.3	1.0
OccFormer [34]	C	12.9	55.7	29.3	30.5	15.5	21.0	22.2	1.1	1.2	1.4	4.6	15.3	4.3	19.2	1.3	2.5	0.6	13.2	4.1	1.8
RenderOcc [29]	C	13.1	57.0	29.2	32.4	**16.6**	19.6	24.8	6.4	2.7	0.2	3.6	26.2	5.8	3.6	0.0	1.1	0.6	9.3	6.1	3.4
LMSCNet [35]	L	14.3	64.1	34.2	29.4	6.3	39.2	27.4	2.2	0.0	0.3	2.2	19.3	3.2	19.4	0.3	0.0	0.1	15.3	5.1	3.5
JS3C-Net [21]	L	22.8	64.2	38.4	34.6	13.3	39.5	34.6	7.9	15.2	8.0	12.2	42.3	18.2	39.3	6.1	5.1	0.5	30.7	17.3	5.2
SSC-RS [44]	L	24.2	71.4	43.8	41.2	11.2	**44.3**	36.1	5.4	13.2	4.5	**14.2**	43.5	25.2	43.1	2.4	1.5	0.4	36.7	16.2	6.1
VPNnet [45]	L	24.8	72.9	44.0	40.2	14.5	44.1	37.1	4.2	13.8	9.3	8.1	45.2	30.2	42.3	3.3	1.8	2.0	32.5	17.8	8.4
OccRWKV [46]	L	25.0	73.1	44.5	40.1	15.9	42.6	36.1	7.1	13.8	7.5	9.8	42.9	30.4	43.1	4.8	1.4	1.2	31.3	18.6	10.6
VoxDet-L [47]	L	25.9	72.8	42.7	37.6	10.1	44.2	37.6	6.4	9.8	6.3	10.2	**45.8**	30.6	**43.5**	2.8	3.4	1.2	32.1	25.6	**29.5**
M-CONet [23]	C&L	20.5	61.6	38.3	29.2	13.2	38.3	33.5	4.3	3.1	2.4	5.5	41.7	20.2	35.4	0.7	2.2	0.4	26.1	18.3	15.6
Co-Occ [22]	C&L	23.9	71.6	41.8	**42.2**	10.3	35.4	39.8	5.7	4.2	3.2	7.9	41.1	30.4	39.8	1.2	3.1	0.3	31.4	**25.7**	19.7
TACOcc (Ours)	C&L	**26.3**	**73.3**	**45.0**	36.5	10.3	40.1	**40.8**	**8.2**	**15.3**	**9.5**	10.8	36.0	**30.8**	40.9	**6.3**	**5.4**	**2.2**	**36.9**	22.3	28.7

TABLE III

ABLATION OF KEY COMPONENTS.

Base	Fuser				Renderer			IoU	mIoU
	Fix k	Dyn.k	Bi.k	MSPF	DGA	$\mathcal{L}_{\text{rgb}}$	$\mathcal{L}_{\text{pc}}$		
✓								34.3	20.7
✓	✓							36.1	22.8
✓		✓						38.2	24.9
✓			✓					39.9	25.2
✓			✓	✓				40.9	25.6
✓			✓	✓	✓			41.2	25.9
✓			✓	✓	✓	✓		42.8	27.2
✓			✓	✓	✓	✓	✓	**44.2**	**28.9**

TABLE IV

FUSER PARAMETER SELECTION.

k	IoU↑	mIoU↑	Latency(s)
1	35.4	22.1	0.52
2	38.1	24.9	0.59
3	41.2	26.1	0.65
1 - 2	40.0	25.2	0.55
1 - 3	**44.2**	**28.9**	0.60

TABLE V

k FREQUENCIES IN SMALL VS. BIG TARGET SCENARIOS

More Small Targets				**More Big Targets**			
k	1	2	3	k	1	2	3
Freq.	41261	63814	31514	Freq.	19539	50432	45624
Prob.	**30.2%**	46.7%	23.1%	Prob.	16.9%	43.6%	**39.4%**

SemanticKITTI, with particular benefits for small objects and long-range regions. Future work includes integrating self-calibration for noisy extrinsics, leveraging radar/events for adverse conditions, and exploring lighter temporal modules for real-time deployment.

TABLE VI

RESULTS WITH DIFFERENT RANGES.

Method	IoU			mIoU		
	25	50	100	25	50	100
M-CONet	60.7	51.1	29.0	36.7	31.4	23.6
Co-Occ	62.3	52.9	41.0	39.3	34.2	26.6
TACOcc	65.8	56.9	44.2	42.8	38.5	28.9
Improvements (%)	**+3.5**	**+4.0**	**+3.2**	**+3.5**	**+4.3**	**+2.8**

TABLE VII

EFFICIENCY ANALYSIS.

Image Size	2D backbone	IoU	mIoU	Memory (G)	Latency (s)
256 × 704	R50	38.9	24.8	10.78	0.47
896 × 1600	R50	40.3	25.9	11.69	0.55
896 × 1600	R101	**44.2**	**28.9**	11.78	0.60

REFERENCES

[1] 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.

[2] 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.

[3] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.

[4] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 400–418.

[5] I. Shepel, V. Adeshkin, I. Belkin, and D. A. Yudin, "Occupancy grid generation with dynamic obstacle segmentation in stereo images," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 779–14 789, 2021.

[6] X. Tan, W. Wu, Z. Zhang, C. Fan, Y. Peng, Z. Zhang, Y. Xie, and L. Ma, "Geocc: Geometrically enhanced 3d occupancy network with implicit-explicit depth fusion and contextual self-supervision," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[7] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin, *et al.*, "Scene as occupancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8406–8415.

[8] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.

[9] H. Xu, J. Chen, S. Meng, Y. Wang, and L.-P. Chau, "A survey on occupancy perception for autonomous driving: The information fusion perspective," *Information Fusion*, vol. 114, p. 102671, 2025.

[10] Y. Zhang, J. Zhang, Z. Wang, J. Xu, and D. Huang, "Vision-based 3d occupancy prediction in autonomous driving: a review and outlook," *arXiv preprint arXiv:2405.02595*, 2024.

[11] Y. Lei, Z. Wang, F. Chen, G. Wang, P. Wang, and Y. Yang, "Recent advances in multi-modal 3d scene understanding: A comprehensive survey and evaluation," *arXiv preprint arXiv:2310.15676*, 2023.

[12] Z. Ming, J. S. Berrio, M. Shan, and S. Worrall, "Inverse++: Vision-centric 3d semantic occupancy prediction assisted with 3d object detection," *arXiv preprint arXiv:2504.04732*, 2025.

[13] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Memory based fusion for multi-modal deep learning," *Information Fusion*, vol. 67, pp. 136–146, 2021.

[14] S. Zhang, Y. Zhai, J. Mei, and Y. Hu, "Fusionocc: Multi-modal fusion for 3d occupancy prediction," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 787–796.

[15] M. Chen, W. Chen, M. Yang, Y. Zhang, T. Han, X. Li, Y. Li, and H. Zhao, "Tgp: Two-modal occupancy prediction with 3d gaussian and sparse points for 3d environment awareness," *arXiv preprint arXiv:2503.09941*, 2025.

[16] 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.

[17] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[18] S. Liu, Y. Hu, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li, "See and think: Disentangling semantic scene completion," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[19] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1746–1754.

[20] 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.

[21] 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*, no. 4, 2021, pp. 3101–3109.

[22] 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.

[23] 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.

[24] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. J. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser, "Panoptic neural fields: A semantic object-aware neural scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 871–12 881.

[25] Z. Wu, T. Liu, L. Luo, Z. Zhong, J. Chen, H. Xiao, C. Hou, H. Lou, Y. Chen, R. Yang, *et al.*, "Mars: An instance-aware, modular and realistic simulator for autonomous driving," in *CAAI International Conference on Artificial Intelligence*. Springer, 2023, pp. 3–15.

[26] W. Gan, N. Mo, H. Xu, and N. Yokoya, "A simple attempt for 3d occupancy estimation in autonomous driving," *CoRR*, 2023.

[27] F. Wimbauer, N. Yang, C. Rupprecht, and D. Cremers, "Behind the scenes: Density fields for single view reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9076–9086.

[28] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[29] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, H. Xie, B. Wang, L. Liu, and S. Zhang, "Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 404–12 411.

[30] M. Pan, L. Liu, J. Liu, P. Huang, L. Wang, S. Zhang, S. Xu, Z. Lai, and K. Yang, "Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering," *arXiv preprint arXiv:2306.09117*, 2023.

[31] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[32] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin, "Understanding straight-through estimator in training activation quantized neural nets," *arXiv preprint arXiv:1903.05662*, 2019.

[33] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arxiv 2022," *arXiv preprint arXiv:2203.17270*, 2022.

[34] 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.

[35] 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.

[36] Y. Ma, J. Mei, X. Yang, L. Wen, W. Xu, J. Zhang, X. Zuo, B. Shi, and Y. Liu, "Licrocc: Teach radar for accurate semantic occupancy prediction using lidar and camera," *IEEE Robotics and Automation Letters*, 2024.

[37] G. Kopanas, T. Leimkühler, G. Rainer, C. Jambon, and G. Drettakis, "Neural point catacaustics for novel-view synthesis of reflections," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–15, 2022.

[38] Y. Fu, S. Liu, A. Kulkarni, J. Kautz, A. A. Efros, and X. Wang, "Colmap-free 3d gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 796–20 805.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[44] J. Mei, Y. Yang, M. Wang, T. Huang, X. Yang, and Y. Liu, "Ssc-rs: Elevate lidar semantic scene completion with representation separation and bev fusion," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1–8.

[45] L. Wang, D. Lin, K. Yang, R. Liu, Q. Guo, W. Xie, M. Wang, L. Liang, Y. Wang, and P. Li, "Voxel proposal network via multi-frame knowledge distillation for semantic scene completion," *Advances in Neural Information Processing Systems*, vol. 37, pp. 101 096–101 115, 2024.

[46] J. Wang, W. Yin, X. Long, X. Zhang, Z. Xing, X. Guo, and Q. Zhang, "Occrwkv: Rethinking efficient 3d semantic occupancy prediction with linear complexity," *arXiv preprint arXiv:2409.19987*, 2024.

[47] W. Li, Z. Yu, and A. Alahi, "Voxdet: Rethinking 3d semantic occupancy prediction as dense object detection," *arXiv preprint arXiv:2506.04623*, 2025.