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

ST-GS: Vision-Based 3D Semantic Occupancy Prediction with Spatial-Temporal Gaussian Splatting

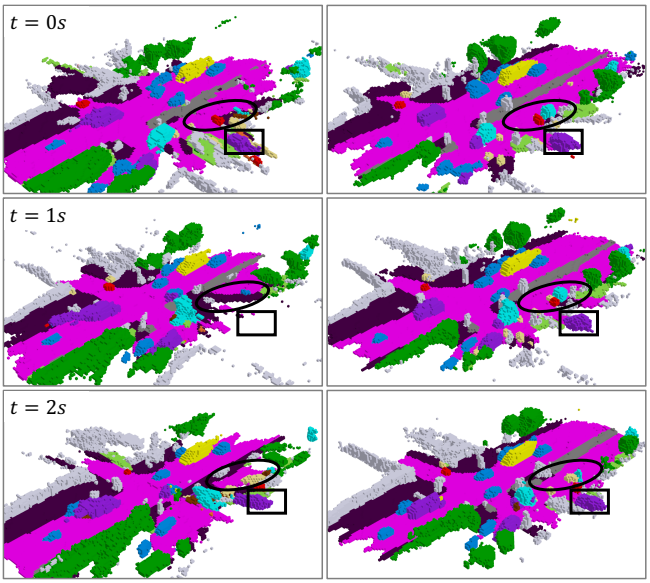
Xiaoyang Yan*, Muleilan Pei*,†, and Shaojie Shen

***Abstract*— 3D occupancy prediction is critical for comprehensive scene understanding in vision-centric autonomous driving. Recent advances have explored utilizing 3D semantic Gaussians to model occupancy while reducing computational overhead, but they remain constrained by insufficient multi-view spatial interaction and limited multi-frame temporal consistency. To overcome these issues, in this paper, we propose a novel Spatial-Temporal Gaussian Splatting (ST-GS) framework to enhance both spatial and temporal modeling in existing Gaussian-based pipelines. Specifically, we develop a guidance-informed spatial aggregation strategy within a dual-mode attention mechanism to strengthen spatial interaction in Gaussian representations. Furthermore, we introduce a geometry-aware temporal fusion scheme that effectively leverages historical context to improve temporal continuity in scene completion. Extensive experiments on the large-scale nuScenes occupancy prediction benchmark showcase that our proposed approach not only achieves state-of-the-art performance but also delivers markedly better temporal consistency compared to existing Gaussian-based methods.**

I. Introduction

Comprehensive 3D scene understanding is a fundamental requirement for modern autonomous driving systems. Vision-based methods have gained increasing attention due to their cost-effectiveness and scalability compared to LiDAR-based approaches [1]. However, accurately modeling complex and irregular objects in dynamic driving scenes remains challenging for mapless driving [2]. Semantic occupancy prediction provides a promising solution by jointly estimating volumetric occupancy and semantic labels of arbitrary-shaped objects in 3D space, thereby improving the reliability and safety of autonomous driving in complex environments [3], [4].

Existing vision-based 3D semantic occupancy prediction methods can be broadly grouped into voxel-based representations, Bird's-Eye View (BEV) projections, and Gaussian-based scene modeling. Voxel-based approaches [5], [6] discretize the 3D space into regular grids and predict semantic occupancy for each voxel, but their cubic complexity leads to high memory usage and limited resolution. BEV-oriented methods [7], [8] project image features into a top-down view for efficient reasoning, yet inevitably discard fine-grained vertical structures that are crucial for dense 3D reconstruction. Recently, Gaussian-based representations [9], [10] have emerged as a promising alternative, using 3D Gaussians as compact primitives to capture continuous geometry while maintaining highly efficient rendering.



(a) Baseline (b) Ours

Fig. 1. Illustration of temporal inconsistency in occupancy prediction. In this example, the side camera views of the ego vehicle are heavily occluded by surrounding vehicles. The baseline method (GaussianFormer [9]) fails to track the identical truck (highlighted by the box) and produces discontinuous drivable surface predictions (highlighted by the ellipse) across frames. In contrast, our proposed ST-GS effectively integrates historical information, delivering accurate and consistent semantic occupancy predictions.

Despite their efficiency and flexibility, existing Gaussian-based approaches encounter two key challenges: (i) they lack the structured spatial priors inherent to grid-based methods, making spatial interaction across views less effective, and (ii) they struggle to maintain temporal consistency across frames, limiting robustness in dynamic environments. To overcome these issues, we aim to strengthen both multi-view spatial interaction and multi-frame temporal consistency in current Gaussian-based semantic occupancy prediction pipelines.

Unlike BEV-oriented methods, which employ predefined spatially ordered grid queries that interact with adjacent areas to capture contextual features, 3D Gaussian primitives are spatially independent and lack inherent neighborhood relations. Consequently, Gaussian-based models rely heavily on reference point sampling strategies to extract spatial information from multi-view images. To address this limitation, we introduce a guidance-informed spatial aggregation strategy built on a dual-mode attention mechanism. Specifically, a Gaussian-guided attention module preserves each primitive's

*Equal Contribution. †Corresponding Author & Project Lead.
This work was supported in part by the Hong Kong Ph.D. Fellowship Scheme, in part by the HKUST Postgraduate Studentship, and in part by the HKUST-DJI Joint Innovation Laboratory.
All authors are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. Email: {xyanaq,mpei,eeshaojie}@ust.hk

ellipsoidal spatial distribution, while a view-guided attention module aggregates complementary spatial and semantic cues from different perspectives. The reference points of these two attention branches are adaptively fused via an efficient yet effective gated feature aggregation network, producing more robust and spatially aligned Gaussian representations.

Furthermore, as incomplete observations in dynamic driving environments degrade temporal coherence and semantic stability, the existing Gaussian-based method [9] often delivers noticeable temporal inconsistency, as shown in Fig. 1(a). To mitigate this issue, we propose a geometry-aware temporal fusion scheme that maintains multi-frame Gaussian primitives and incorporates relevant historical information into the current scene representations while explicitly accounting for geometric correspondences. This is achieved by employing a lightweight gated temporal feature fusion module, which significantly enhances temporal consistency across frames in semantic scene completion, as illustrated in Fig. 1(b).

In summary, our work makes the following contributions:

- We introduce Spatial-Temporal Gaussian Splatting (ST-GS), a novel framework that effectively improves multi-view spatial interaction and multi-frame temporal consistency for Gaussian-based occupancy prediction.
- We propose a guidance-informed spatial aggregation strategy within a dual-mode attention mechanism to enhance spatial modeling of 3D Gaussian representations.
- We design a geometry-aware temporal fusion scheme with a gated feature fusion module to integrate historical contexts while preserving geometric correspondences.
- Our ST-GS achieves state-of-the-art performance on the nuScenes dataset, and further exhibits superior temporal consistency relative to existing Gaussian-based models.

II. Related Work

A. 3D Semantic Occupancy Prediction

3D semantic occupancy prediction has gained significant attention in recent years, as it provides a more comprehensive representation of surrounding environments compared to conventional 3D detection or segmentation tasks [11]. This capability makes it particularly vital for autonomous driving [12]. Early studies primarily relied on LiDAR point clouds, leveraging their precise geometric information for occupancy estimation [13], [14]. However, LiDAR sensors are expensive and often degrade under adverse weather or poor lighting conditions. To overcome these limitations, vision-based 3D semantic occupancy prediction has emerged as an active area of research. Recent works have adopted voxel-based representations to model 3D occupancy [15], [16], [17], [18], as voxels effectively capture continuous 3D structures within a defined spatial volume. Nonetheless, the inherent sparsity of real-world 3D scenes makes voxel-based methods computationally expensive. Consequently, more efficient scene representations have been explored. TPVFormer [19] employs a tri-plane representation, which enforces stronger spatial constraints compared to BEV-based approaches [7] that project the scene onto a single plane. Despite these advances, both voxel-based and BEV-based methods require a dense grid of representations to model the 3D environment. In contrast, Gaussian-based approaches encode 3D scenes more compactly, using fewer primitives while maintaining geometric fidelity. Motivated by these advantages, our work focuses on Gaussian-based semantic occupancy prediction.

B. Gaussian-Based Scene Representation

3D Gaussian representations have been widely adopted in scene reconstruction [20] due to their strong modeling capabilities and compact encoding of 3D structures. In contrast to dense voxel grids, Gaussian primitives represent 3D semantic occupancy as anisotropic elements that naturally adapt to scene complexity: more primitives are allocated to regions with rich geometry and semantics, while textureless areas are represented with far fewer. Building on these properties, GaussianFormer [9] employs sparse 3D Gaussian primitives and leverages Gaussian-to-voxel splatting to predict semantic occupancy, enabling better adaptability to complicated driving environments than grid-based methods. Furthermore, GaussianFormer-2 [10] enhances both efficiency and accuracy through introducing a probabilistic distribution modeling strategy to optimize Gaussian initialization, thereby alleviating inefficiencies caused by large empty regions in 3D space. Nevertheless, existing Gaussian-based approaches provide limited treatment of geometric priors and historical information. This motivates our work to strengthen spatial-temporal modeling within Gaussian-based frameworks.

C. Spatial-Temporal Modeling

Driving scenes are highly dynamic and often suffer from severe occlusions and fast-moving objects, which make reliable 3D semantic occupancy prediction particularly challenging. Several approaches exploit spatial-temporal information to enhance spatial reasoning and improve prediction robustness. PanoOcc [21] extracts multi-frame image features to construct a set of voxel queries, which are subsequently aligned in unified spatial space and fused through a dedicated temporal encoder to produce a unified occupancy representation. BEVFormer [7] maintains BEV queries across multiple frames and employs a temporal self-attention mechanism to effectively exchange cross-frame features, thereby enhancing scene understanding. GaussianWorld [22] employs a world-model-driven framework to process video streams to exploit historical observations. ST-Occ [23] improves the temporal dependency by utilizing spatiotemporal memory across multiple frames. However, current Gaussian-based methods still fall short in spatial-temporal modeling, and thus, we intend to further alleviate this by fully leveraging geometric spatial priors and effectively integrating historical information to boost both prediction accuracy and temporal consistency.

III. Methodology

A. Gaussian-Based Occupancy Prediction

The 3D semantic occupancy prediction task aims to estimate the volumetric occupancy and semantic labels of each voxel in 3D space. The task takes N images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ as

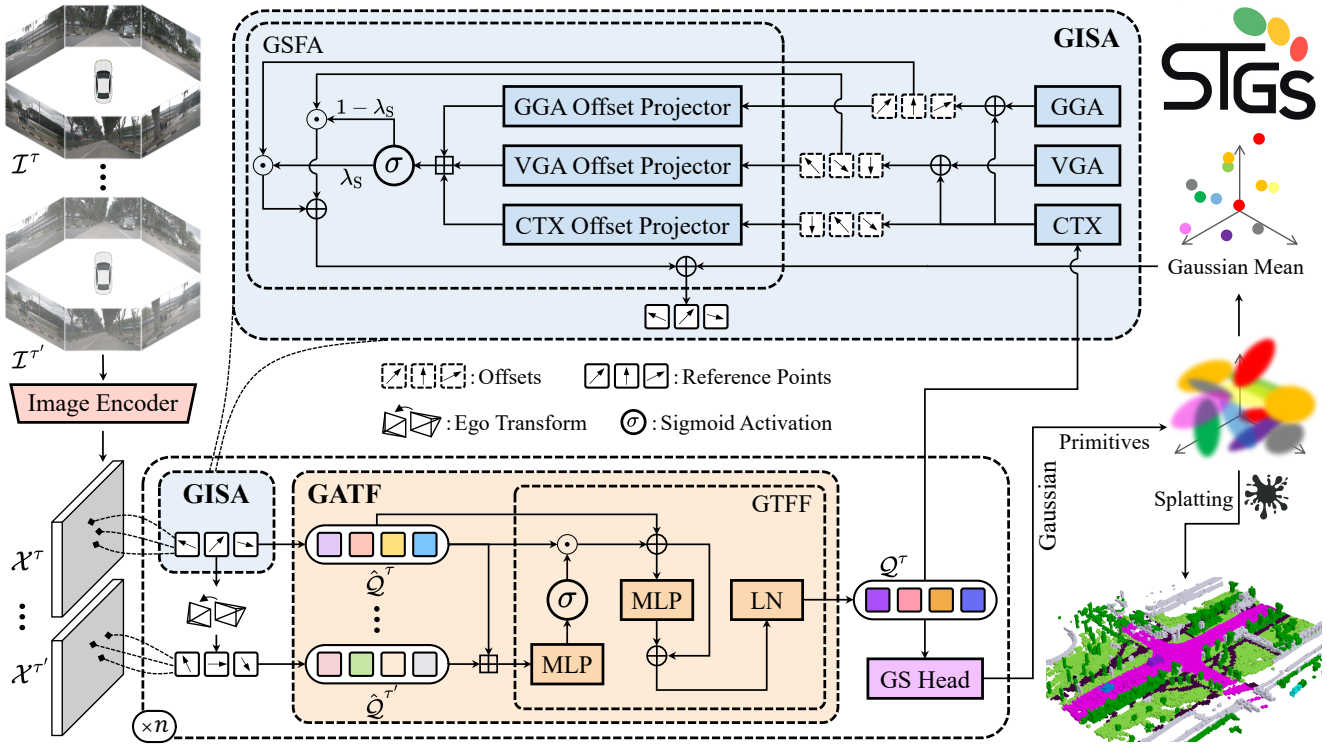


Fig. 2. Overview of our ST-GS architecture, demonstrating how it enhances the existing Gaussian-based occupancy prediction model in multi-view spatial interaction and multi-frame temporal consistency.

input and predicts a dense voxel-based semantic occupancy $\mathcal{O} \in C^{X \times Y \times Z}$, where C is the set of semantic classes. Each voxel is associated with a semantic category. Rather than directly predicting voxel-wise occupancy, the 3D Gaussian representation models a scene as a set of K learnable 3D Gaussian primitives $\mathcal{G} = \{\mathrm{G}_i\}_{i=1}^{K}$. Each primitive G_i is defined by its center $\mathbf{m}_i \in \mathbb{R}^3$, scale $\mathbf{s}_i \in \mathbb{R}^3$, rotation vectors $\mathbf{r}_i \in \mathbb{R}^4$, opacity $\alpha_i \in \mathbb{R}^1$, and the semantic logits $\mathbf{c}_i \in \mathbb{R}^{|C|}$ corresponding to $|C|$ categories, respectively. Moreover, we adopt the Gaussian-to-voxel splatting scheme [9] to render Gaussian primitives $\mathcal{G}$ into the voxel space. For each voxel center $\mathbf{x}$, its semantic value can be calculated by:

$$\mathcal{O}(\mathbf{x}) = \sum_{i=1}^{K} \alpha_i \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m}_i)\right)\mathbf{c}_i, \quad (1)$$

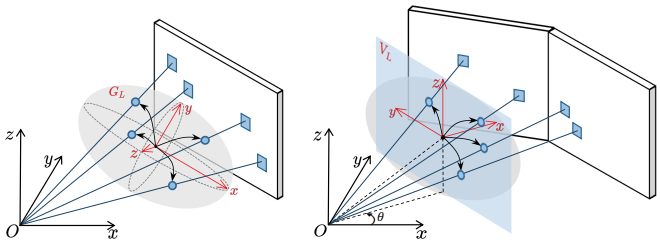
where $\Sigma = RSS^\top R^\top$ denotes the covariance matrix. Herein, $S = \mathrm{diag}(\mathbf{s})$ represents a diagonal matrix whose diagonal entries are the scale components (s_x, s_y, s_z), and $R = \mathrm{q2r}(\mathbf{r})$ denotes the 3×3 rotation matrix obtained by converting the quaternion $\mathbf{r}$ through the $\mathrm{q2r}(\cdot)$ operation.

B. Framework Overview

The overall pipeline of our proposed ST-GS framework is illustrated in Fig. 2, which enhances the Gaussian-based occupancy prediction paradigm by effectively incorporating spatial-temporal information. Given a sequence of τ consecutive surround-view images $\{\mathcal{I}^t\}_{t=1}^{\tau}$, we first extract multi-view 2D features $\{\mathcal{X}^t\}_{t=1}^{\tau}$ using a shared image encoder and maintain a set of 3D Gaussian embeddings $\{\mathcal{Q}^t\}_{t=1}^{\tau}$, which act as learnable queries that adaptively sample and aggregate image features to construct 3D representations. Further, the Guidance-Informed Spatial Aggregation (GISA) strategy is introduced to bridge 2D visual features and 3D Gaussian embeddings through a dual-mode attention mechanism: the Gaussian-Guided Attention (GGA) that exploits the intrinsic 3D Gaussian attributes to refine local feature sampling, and the View-Guided Attention (VGA) that leverages spatial-semantic continuity across multi-view images by adaptively sampling along the camera rays. An efficient Gated Spatial Feature Aggregation (GSFA) module is subsequently applied to yield the final reference points. To further improve temporal coherence, the Geometry-Aware Temporal Fusion (GATF) scheme is designed to explicitly align Gaussian embeddings across frames using ego-motion transformations and selectively integrate relevant historical information into the current keyframe representation through an efficient Gated Temporal Feature Fusion (GTFF) module. Finally, the enhanced Gaussian embeddings are decoded into Gaussian primitives by a lightweight GS head [24], which performs Gaussian-to-voxel splatting to generate dense semantic occupancy voxels.

C. Guidance-Informed Spatial Aggregation

To fully exploit spatial priors from camera viewpoints, we propose the Guidance-Informed Spatial Aggregation (GISA) strategy, which bridges 2D visual features and 3D Gaussian embeddings by dynamically determining how embeddings attend to and query relevant information from the image feature space. GISA employs a dual-mode attention mechanism to incorporate two complementary reference points, enabling



(a) GGA (b) VGA

Fig. 3. Feature sampling paradigms of offsets for GGA and VGA.

more effective spatial feature sampling. Formally, given the 2D image feature maps $\mathcal{X}$ extracted from multi-view cameras and the reference points $\mathcal{P}$ that aggregate offsets from both Gaussian-guided and view-guided attention mechanisms, the single-frame Gaussian embedding $\mathcal{Q} = \{\mathcal{Q}_i\}_{i=1}^{K} \in \mathbb{R}^{K\times\mathcal{D}}$, with $\mathcal{D}$ denoting the channel dimension, is updated to $\hat{\mathcal{Q}}$ via the following deformable cross-attention operation:

$$\hat{\mathcal{Q}} = \texttt{DeformAttn}(\mathcal{Q}, \mathcal{X}, \mathcal{P}_{2D}), \tag{2}$$

$$\mathcal{P}_{2D} = \texttt{Warp}(\mathcal{P}, \mathcal{K}^{\text{cam}}, \mathcal{T}^{\text{cam}}), \tag{3}$$

where $\texttt{DeformAttn}(\cdot)$ is a deformable cross-attention operation, and $\texttt{Warp}(\cdot)$ denotes the warping function that projects 3D reference points $\mathcal{P}$ onto the 2D image plane, yielding $\mathcal{P}_{2D}$, using the camera intrinsics $\mathcal{K}^{\text{cam}}$ and extrinsics $\mathcal{T}^{\text{cam}}$.

1) Gaussian-Guided Attention: The Gaussian-Guided Attention (GGA) mechanism generates adaptive sampling offsets directly from the parameters of each Gaussian, leveraging their intrinsic encoding of the scene's structural attributes. As shown in Fig. 3(a), GGA uses the Gaussian mean and covariance as geometric guidance to adaptively produce offsets aligned with each Gaussian ellipsoidal distribution. Further, for each Gaussian instance, we initialize the sampling offsets $\mathcal{P}^{G_L} = \{\mathcal{P}_i^{G_L} \in \mathbb{R}^3\}_{i=1}^{M}$ consisting of M points structured by 3D grid, with offsets defined in local Gaussian coordinate frame G_L. The scaled offset proposal is then combined with learned offsets to form local sampling offsets $\Delta\mathcal{P}^{G_L}$:

$$\Delta\mathcal{P}^{G_L} = s^G\mathcal{P}^{G_L} + \Phi_\Delta(\mathcal{Q}_i), \tag{4}$$

where s^G is a learnable scaling factor regulating the sampling radius in the Gaussian coordinate space, and $\Phi_\Delta(\cdot)$ is the sampling offset predictor.

To transform the local offsets into the perception coordinate frame where each Gaussian is defined, we apply the corresponding rotation R^G and scale S^G transformations to obtain the final offset $\Delta\mathcal{P}^G$:

$$\Delta\mathcal{P}^G = R^G S^G \Delta\mathcal{P}^{G_L}. \tag{5}$$

2) View-Guided Attention: Unlike GGA, which utilizes a predefined uniform sampling paradigm consisting of a set of directions evenly distributed in 3D space, motivated by view attention [25], we design a View-Guided Attention (VGA) mechanism. As demonstrated in Fig. 3(b), VGA generates offsets along the camera viewing directions, enabling more effective spatial information aggregation across overlapping multi-view image features by leveraging cross-view geometric priors. Similarly, for each Gaussian instance, we initialize a set of sampling offsets $\mathcal{P}^{V_L} = \{\mathcal{P}_i^{V_L} \in \mathbb{R}^3\}_{i=1}^{M}$ based on a 2D grid in the local view coordinate frame V_L (y–z plane). To achieve instance-specific adaptivity, we also predict an offset for each Gaussian instance from its embedding. The local sampling offsets $\Delta\mathcal{P}^{V_L}$ can then be obtained by:

$$\Delta\mathcal{P}^{V_L} = s^V\mathcal{P}^{V_L} + \Phi_\Delta(\mathcal{Q}_i), \tag{6}$$

where s^V is a learnable scalar parameter. Next, we transform the V_L coordinate frame to the perception coordinate frame by computing a rotation matrix $R^V(\theta)$ based on the azimuth angle θ of the Gaussian center $\mathbf{m}_i$, i.e.,

$$R^V(\theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{7}$$

The final VGA offset $\Delta\mathcal{P}^V$ can be obtained by:

$$\Delta\mathcal{P}^V = R^V(\theta)\Delta\mathcal{P}^{V_L}. \tag{8}$$

3) Gated Spatial Feature Aggregation: To better leverage the advantages of both attention mechanisms, we introduce the Gated Spatial Feature Aggregation (GSFA) module to effectively integrate the offsets from GGA and VGA. GSFA employs an attention-driven gating paradigm to dynamically balance the contributions of GGA and VGA.

Given $\Delta\mathcal{P}^G$, $\Delta\mathcal{P}^V$, and the context-aware offset $\Delta\mathcal{P}^{\text{ctx}} = \Phi_\Delta(\mathcal{Q}_i)$, we first project them into a latent space, obtaining the corresponding embeddings $\mathcal{F}_G$, $\mathcal{F}_V$, and $\mathcal{F}_{\text{ctx}}$, respectively. These embeddings are then concatenated and passed through a sigmoid activation function $\sigma(\cdot)$ to generate an adaptive gate $\lambda_{\text{S}} \in [0,1]^{K\times M}$:

$$\lambda_{\text{S}} = \sigma(\mathcal{F}_G \boxplus \mathcal{F}_V \boxplus \mathcal{F}_{\text{ctx}}), \tag{9}$$

where $\boxplus$ denotes concatenation along the feature dimension. The final aggregated offset is derived by:

$$\Delta\mathcal{P} = \lambda_{\text{S}} \odot \Delta\mathcal{P}^G + (1-\lambda_{\text{S}}) \odot \Delta\mathcal{P}^V, \tag{10}$$

where $\odot$ denotes element-wise multiplication. Consequently, the reference points $\mathcal{P}$ are obtained by adding the aggregated offset $\Delta\mathcal{P}$ to the Gaussian centers $\mathbf{m}_i$:

$$\mathcal{P} = \mathbf{m}_i + \Delta\mathcal{P}. \tag{11}$$

Through GSFA, the dual-mode reference points are seamlessly fused, yielding informative spatial feature attributes.

D. Geometry-Aware Temporal Fusion

To fully exploit streaming contexts in autonomous driving scenarios, we introduce a Geometry-Aware Temporal Fusion (GATF) scheme that incorporates historical information to enhance the Gaussian representation capability of the current frame. GATF operates on Gaussian embeddings produced by GISA and is designed to model temporal dependencies. By explicitly leveraging ego-motion to establish geometric correspondence across frames and selectively aggregating relevant historical information, GATF significantly enhances multi-frame feature alignment and temporal consistency.

1) Inter-frame Geometric Correspondence: Accurate geometric correspondence between multiple frames is a prerequisite for effective temporal fusion. Since Gaussian embeddings from different timesteps originate from asynchronous observations, the associated reference points are often temporally misaligned. To address this, we explicitly align the reference points of historical frames with those of the current frame, thereby ensuring geometric consistency.

Formally, the reference points $\mathcal{P}^{\tau}$ of the current frame τ are transformed into the coordinate system of a historical frame $\tau' \in [1, \tau-1]$ through ego-motion information:

$$\mathcal{P}^{\tau'} = \mathcal{T}^{\tau \to \tau'} \mathcal{P}^{\tau}, \tag{12}$$

where $\mathcal{P}^{\tau'}$ denotes the aligned reference points in the historical frame, and $\mathcal{T}^{\tau \to \tau'}$ is the rigid-body transformation from the current frame τ to the historical frame τ'.

2) Gated Temporal Feature Fusion: Once geometric correspondence is established, the GISA-updated multi-frame Gaussian embeddings $Q = \{\hat{\mathcal{Q}}^t\}_{t=1}^{\tau} \in \mathbb{R}^{\tau \times K \times \mathcal{D}}$ become available for temporal fusion. The objective is to selectively integrate historical information into the current embedding $\hat{\mathcal{Q}}^{\tau}$, while effectively suppressing inconsistent features arising from occlusions or dynamic objects.

To this end, we introduce a lightweight Gated Temporal Feature Fusion (GTFF) module that adaptively incorporates historical Gaussian embeddings into the current frame. The module first predicts an adaptive fusion gate $\lambda_{\mathrm{T}} \in [0,1]^{K \times \mathcal{D}}$ via a temporal weight generator:

$$\lambda_{\mathrm{T}} = \sigma\big(\mathtt{MLP}(Q)\big), \tag{13}$$

where $\mathtt{MLP}(\cdot)$ represents a multi-layer perceptron block.

Next, the gate λ_{T} modulates the contribution of historical embeddings relative to the current frame, producing a gated embedding $\tilde{\mathcal{Q}}^{\tau}$:

$$\tilde{\mathcal{Q}}^{\tau} = \hat{\mathcal{Q}}^{\tau} + \lambda_{\mathrm{T}} \odot \hat{\mathcal{Q}}^{\tau}. \tag{14}$$

Consequently, the final Gaussian embedding of the current frame, $\mathcal{Q}^{\tau} \in \mathbb{R}^{K \times \mathcal{D}}$, is obtained via a residual refinement:

$$\mathcal{Q}^{\tau} = \mathtt{LN}\Big(\hat{\mathcal{Q}}^{\tau} + \mathtt{MLP}(\tilde{\mathcal{Q}}^{\tau})\Big), \tag{15}$$

where $\mathtt{LN}(\cdot)$ denotes layer normalization.

By jointly learning adaptive temporal weights and residual refinement, the GTFF module effectively integrates historical information, strengthening multi-frame temporal consistency.

E. Training Loss

After the Gaussian-to-voxel splatting process, we obtain the semantic occupancy prediction. Consistent with [9], [19], we adopt the cross entropy loss and the Lovász-Softmax loss [26] to optimize the output of each block.

IV. Experiments and Results

A. Experimental Setup

1) Dataset: We evaluate our approach on the nuScenes dataset [28], which consists of 1,000 driving sequences, each lasting 20 seconds. The dataset provides annotations at 2 Hz and includes six synchronized cameras covering a full 360° horizontal field of view. Following the established protocol [9], [10], we adopt the official split of 700 scenes for training and 150 for validation. For supervision, we use the semantic occupancy ground truth provided by SurroundOcc [15]. The annotated space spans a volume of $100m \times 100m \times 8m$, centered on the ego vehicle. The target occupancy representation is discretized into a $200 \times 200 \times 16$ voxel grid, where each voxel corresponds to $0.5m$ in physical space. This covers spatial ranges of $[-50m, 50m]$ along both horizontal axes and $[-5m, 3m]$ in height. RGB images from all six cameras are used at their native resolution of 1600×900 pixels.

2) Evaluation Metrics: We assess the performance of 3D semantic occupancy prediction using two standard metrics following [16]. For Scene Completion (SC), we report class-agnostic Intersection-over-Union (IoU), which evaluates geometric accuracy irrespective of semantics. For Semantic Scene Completion (SSC), we report mean IoU (mIoU) across all semantic classes, reflecting both geometric and semantic quality. Formally,

$$\mathrm{IoU} = \frac{\mathrm{TP}^{(c_n)}}{\mathrm{TP}^{(c_n)} + \mathrm{FP}^{(c_n)} + \mathrm{FN}^{(c_n)}}, \tag{16}$$

$$\mathrm{mIoU} = \frac{1}{|C'|} \sum_{i \in C'} \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)} + \mathrm{FN}^{(i)}}, \tag{17}$$

where $\mathrm{TP}^{(i)}$, $\mathrm{FP}^{(i)}$, and $\mathrm{FN}^{(i)}$ denote the true positives, false positives, and false negatives, respectively, for class $i \in C'$, and C' is the set of non-empty semantic classes. The variable c_n refers to non-empty classes in the SC evaluation.

Moreover, we employ the Spatial-Temporal Classification Variability (STCV) metric [23] to quantify the temporal consistency of occupancy predictions across consecutive frames. STCV calculates the proportion of classification alterations in non-empty voxels relative to all non-empty voxels over L successive frames within each scene, defined as:

$$\mathrm{STCV} = \frac{1}{\mathrm{L}-1} \sum_{t=1}^{\mathrm{L}-1} \frac{\sum_{\mathcal{V}_t \wedge \mathcal{V}_{t+1}} \mathbb{1}\big[\mathcal{O}^t \neq \mathcal{O}^{t+1}\big]}{\sum_{\mathcal{V}_t \wedge \mathcal{V}_{t+1}} \mathbb{1}[\mathcal{O}^t]}, \tag{18}$$

where $\mathcal{V}_t$ denotes the set of non-empty voxels at the frame t, and $\mathbb{1}[\cdot]$ represents the indicator function. For a comprehensive evaluation of temporal consistency, we report the mean STCV (mSTCV), minimum STCV (minSTCV), and maximum STCV (maxSTCV) across all scenes.

3) Implementation Details: We adopt the default ResNet-101-DCN [29] as the backbone for image feature extraction, consistent with existing Gaussian-based methods [9], [10], to ensure fair comparison. A Feature Pyramid Network (FPN) [30] is employed to capture multi-scale features at downsampling ratios of $\{4\times, 8\times, 16\times, 32\times\}$. The channel dimension $\mathcal{D}$ is fixed to 128 across all models. The Gaussian-based decoder comprises $n = 4$ stacked blocks with $K = 25,600$ Gaussian primitives. We train our model using AdamW [31] with a weight decay of 0.01. The learning rate follows a warm-up strategy for the first 500 iterations, reaching a maximum value of 2×10^{-4}, and then decays following a

TABLE I

3D SEMANTIC OCCUPANCY PREDICTION RESULTS ON THE NUSCENES VALIDATION SPLIT.

Method	Venue	SC IoU	SSC mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [16]	CVPR 2022	23.96	7.31	4.03	0.35	8.00	8.04	2.90	0.28	1.16	0.67	4.01	4.35	27.72	5.20	15.13	11.29	9.03	14.86
Atlas [27]	ECCV 2020	28.66	15.00	10.64	5.68	19.66	24.94	8.90	8.84	6.47	3.28	10.42	16.21	34.86	15.46	21.89	20.95	11.21	20.54
BEVFormer [7]	ECCV 2022	30.50	16.75	14.22	6.58	23.46	28.28	8.66	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer [19]	CVPR 2023	11.51	11.66	16.14	7.17	22.63	17.13	8.83	11.39	10.46	8.23	9.43	17.02	8.07	13.64	13.85	10.34	4.90	7.37
TPVFormer†[19]	CVPR 2023	30.86	17.10	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
OccFormer [17]	ICCV 2023	31.39	19.03	18.65	10.41	23.92	30.29	10.31	14.19	13.59	10.13	12.49	20.77	38.78	19.79	24.19	22.21	13.48	21.35
SurroundOcc [15]	ICCV 2023	31.49	20.30	20.59	11.68	**28.06**	30.86	10.70	15.14	**14.09**	**12.06**	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
GaussianFormer [9]	ECCV 2024	29.83	19.10	19.52	11.26	26.11	29.78	10.47	13.83	12.58	8.67	12.74	21.57	39.63	23.28	24.46	22.99	9.59	19.12
GaussianFormer-2*[10]	CVPR 2025	30.56	20.02	20.15	12.99	27.61	30.23	11.19	15.31	12.64	9.63	13.31	22.26	39.68	23.47	25.62	23.20	12.25	20.73
ST-GS(Ours)	ICRA 2026	**32.88**	**21.43**	**21.04**	**14.13**	27.78	**31.62**	**11.85**	**17.84**	13.63	10.76	**14.85**	**23.22**	**41.88**	**24.40**	**26.71**	**24.70**	**15.00**	**23.48**

† means supervision with dense occupancy annotations [15].
* indicates results of the 128-channel dimension for a fair comparison [10].

TABLE II

EVALUATION OF TEMPORAL CONSISTENCY ON THE NUSCENES DATASET.

Method	mSTCV(%)	minSTCV(%)	maxSTCV(%)
FB-OCC [32]	12.18	-	-
ST-Occ [23]	8.68	-	-
GaussianFormer [9]	6.52	1.59	12.71
GaussianFormer-2 [10]	5.97	1.94	10.86
ST-GS (Ours)	**4.47**	**1.03**	**8.53**

A "-" denotes the lack of relevant data.

TABLE III

ABLATION ON COMPONENTS OF THE GISA STRATEGY.

Method	GGA	VGA	GSFA	IoU	mIoU
GaussianFormer [9]				29.83	19.10
w/ GGA Only	✓			30.91	19.85
w/ VGA Only		✓		30.97	19.92
w/ GISA (Ours)	✓	✓	✓	**31.51**	**20.27**

TABLE IV

EFFECT OF SEQUENCE LENGTH.

# of frames	IoU	mIoU
1	31.51	20.27
2	32.01	20.82
3	**32.88**	**21.43**

TABLE V

EFFECT OF FUSION MODE.

Fusion Mode	IoU	mIoU
Loose	32.11	21.11
Tight	32.44	21.13
Coupled	**32.88**	**21.43**

cosine annealing schedule. To improve model generalization, we leverage standard data augmentations, including random cropping, flipping, resizing, and photometric distortions.

B. Quantitative Results

1) Main Results: Table I offers a comprehensive comparison of our ST-GS with existing methods on the nuScenes validation split, with the best and second-best results highlighted in **bold** and underlined, respectively. Our ST-GS consistently outperforms both previous voxel-based approaches and recent Gaussian-based methods. Compared to our baseline, GaussianFormer [9], ST-GS achieves notable improvements of 10.22% in IoU and 12.20% in mIoU. Even against its successor, GaussianFormer-2 [10], ST-GS delivers substantial gains of 7.59% in IoU and 7.04% in mIoU. Moreover, our method attains the highest performance across most semantic categories. These results demonstrate the effectiveness of our framework and underscore that enhancing spatial interactions and incorporating temporal information are both crucial for accurate 3D semantic occupancy prediction.

2) Temporal Consistency Results: We conduct the evaluation of temporal consistency on the nuScenes validation set. All metrics follow the lower-is-better criterion. As shown in Table II, our ST-GS achieves the best performance across all metrics, surpassing both previous voxel-based approaches and recent Gaussian-based methods. In particular, relative to our baseline, GaussianFormer [9], ST-GS reduces temporal inconsistency by 31.44% in mSTCV, 35.22% in minSTCV, and 32.89% in maxSTCV, respectively. These results highlight that our framework substantially improves multi-frame temporal consistency, leading to more robust and stable 3D semantic occupancy prediction.

C. Ablation Studies

We perform ablation studies on the nuScenes validation split to assess the effectiveness of the proposed GISA strategy and GATF scheme.

1) Effect of the GISA Strategy: In Table III, we present the ablation results for each component of GISA, including GGA, VGA, and GSFA. Compared with the baseline GaussianFormer [9], both GGA and VGA individually yield improvements in IoU and mIoU, validating the contribution of each module. Moreover, when combined through GSFA, performance is further enhanced, highlighting the advantage of the fusion strategy in facilitating multi-view spatial feature interactions and achieving substantial performance gains in 3D semantic occupancy prediction accuracy.

2) Effect of the GATF Scheme: We further evaluate the effectiveness of the proposed GATF scheme. As shown in Table IV, we first investigate the effect of temporal sequence length. The results demonstrate that increasing the number

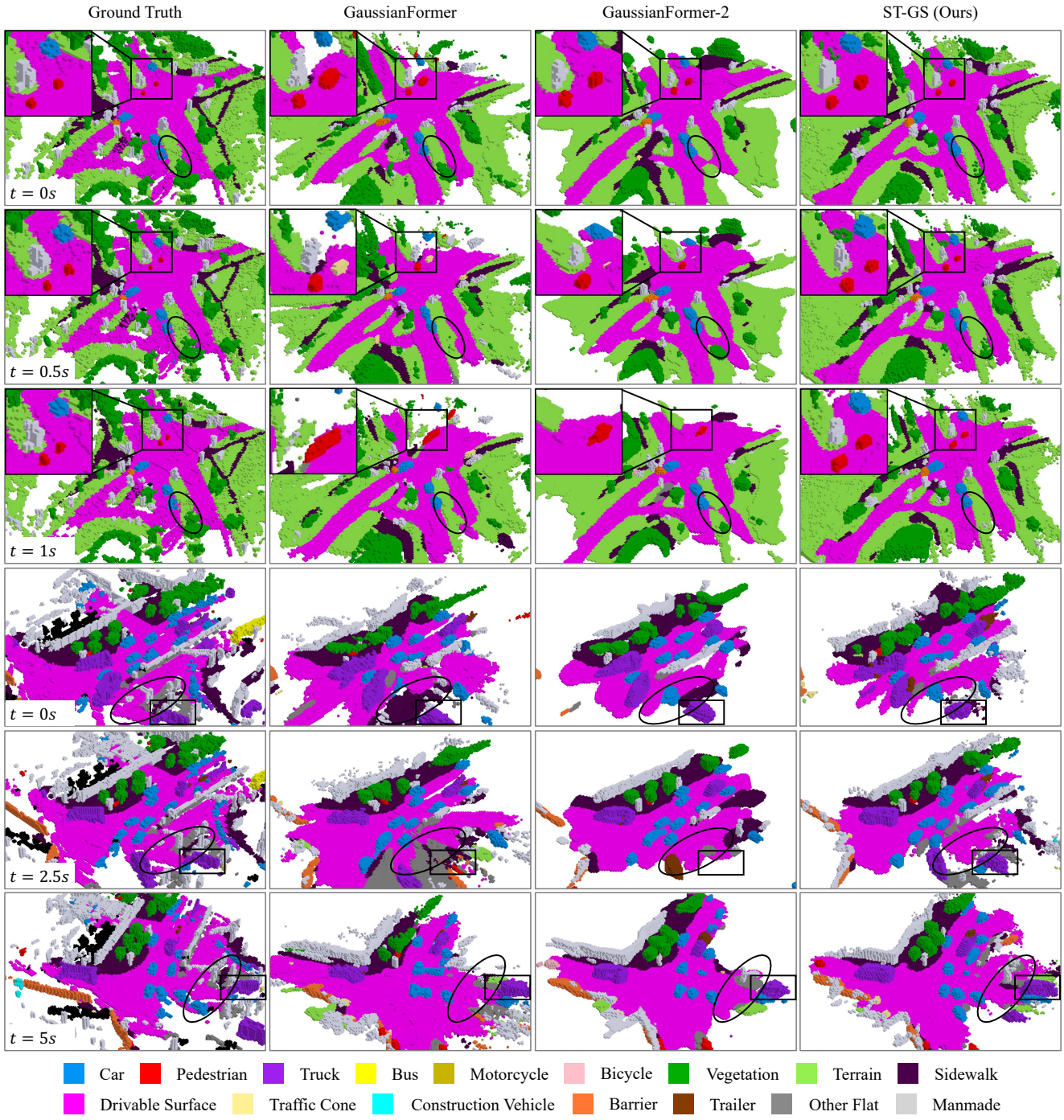


Fig. 4. Qualitative comparison of the baseline GaussianFormer [9], GaussianFormer-2 [10], and our proposed ST-GS. Visualization results of three-timestamp predictions from two distinct driving sequences show that ST-GS delivers more spatially accurate and temporally consistent semantic occupancy predictions.

of frames consistently improves all metrics, confirming that appropriately integrating richer historical context enhances prediction performance, as long as computational overhead remains controlled. We then analyze the impact of different fusion modes of GTFF on prediction accuracy. Specifically, we examine three distinct fusion configurations: (i) loose mode, where GTFF aggregates historical embeddings into the current frame only once at the final stage after the four stacked blocks; (ii) tight mode, where GTFF is applied within each individual block for fine-grained temporal fusion; and (iii) coupled mode, which combines both strategies to attain more comprehensive feature integration. As reported in Table V, all fusion modes enhance prediction performance, with the coupled mode delivering the largest improvements in both IoU and mIoU. These findings underscore that incorporating richer temporal information and adopting effective fusion schemes are essential for achieving significant performance gains in 3D semantic occupancy prediction.

D. Qualitative Results

We present visualizations of two representative scenarios with varying time intervals from the nuScenes validation set to demonstrate the superiority of our ST-GS over the baseline GaussianFormer [9] and GaussianFormer-2 [10]. As shown in the upper group of Fig. 4, all methods correctly predict two pedestrians and one car (box) in the initial frame, whereas GaussianFormer-2 misclassifies the drivable surface (ellipse). In the subsequent frame, GaussianFormer misclassifies one pedestrian category, and GaussianFormer-2 fails to detect one pedestrian. In the final frame, both fail to accurately predict the pedestrians and the car. In contrast, our ST-GS preserves category accuracy and instance-level stability across all three consecutive frames, effectively handling both small objects and large structures. Furthermore, the lower group of Fig. 4 showcases an intersection with heavy occlusion over a longer time span. In this case, both of the other two methods fail to stably track the same truck (box) and generate temporally inconsistent predictions for drivable surfaces (ellipse), whereas our ST-GS accurately and consistently identifies them across frames. Overall, these qualitative results highlight that ST-GS substantially improves both prediction accuracy and temporal consistency, even under long-horizon scenarios. Additional qualitative results are provided in the supplementary video.

V. Conclusion

In this paper, we propose ST-GS, an innovative framework designed to strengthen both spatial and temporal modeling in the Gaussian-based semantic occupancy prediction pipeline. Specifically, ST-GS improves multi-view spatial interaction through the GISA strategy and enforces multi-frame temporal consistency via the GATF scheme. Experimental results on the large-scale nuScenes occupancy prediction benchmark exhibit that ST-GS achieves 32.88 IoU and 21.43 mIoU, outperforming prior voxel-based methods and recent Gaussian-based approaches by substantial margins. Furthermore, ST-GS reduces temporal inconsistency by over 30% in mSTCV relative to the baseline, demonstrating the effectiveness of enhancing spatial interactions and incorporating temporal information in delivering accurate and robust 3D semantic occupancy predictions. Future work will focus on improving the efficiency of the proposed framework by exploring advanced architectures, such as Gamba [33] which integrates Gaussian splatting with the Mamba [34] training paradigm.

References

[1] 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 *ICRA*, 2024.

[2] M. Pei, J. Shan, P. Li, J. Shi, J. Huo, Y. Gao, and S. Shen, "Sept: Standard-definition map enhanced scene perception and topology reasoning for autonomous driving," *IEEE Robotics and Automation Letters*, 2025.

[3] R. Marcuzzi, L. Nunes, E. Marks, L. Wiesmann, T. Läbe, J. Behley, and C. Stachniss, "Sfmocc: Vision-based 3d semantic occupancy prediction in urban environments," *IEEE Robotics and Automation Letters*, 2025.

[4] M. Pei, S. Shi, and S. Shen, "Advancing multi-agent traffic simulation via r1-style reinforcement fine-tuning," in *ICLR*, 2026.

[5] 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 *CVPR*, 2023.

[6] H. Jiang, T. Cheng, N. Gao, *et al.*, "Symphonize 3d semantic scene completion with contextual instance queries," in *CVPR*, 2024.

[7] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, *et al.*, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022.

[8] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, *et al.*, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *AAAI*, 2023.

[9] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction," in *ECCV*, 2024.

[10] Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du, and J. Lu, "Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction," in *CVPR*, 2025.

[11] M. Pei, H. An, B. Liu, and C. Wang, "An improved dyna-q algorithm for mobile robot path planning in unknown dynamic environment," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[12] M. Pei, S. Shi, X. Chen, X. Liu, and S. Shen, "Foresight in motion: Reinforcing trajectory prediction with reward heuristics," in *ICCV*, 2025.

[13] 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 *AAAI*, 2021.

[14] 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 *ICCV*, 2023.

[15] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, *et al.*, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *ICCV*, 2023.

[16] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *CVPR*, 2022.

[17] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *ICCV*, 2023.

[18] M. Pei, S. Shi, L. Zhang, P. Li, and S. Shen, "Goirl: Graph-oriented inverse reinforcement learning for multimodal trajectory prediction," in *ICML*, 2025.

[19] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, *et al.*, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *CVPR*, 2023.

[20] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, 2023.

[21] Y. Wang, Y. Chen, *et al.*, "Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation," in *CVPR*, 2024.

[22] S. Zuo, W. Zheng, Y. Huang, J. Zhou, *et al.*, "Gaussianworld: Gaussian world model for streaming 3d occupancy prediction," in *CVPR*, 2025.

[23] Z. Leng, J. Yang, W. Yi, and B. Zhou, "Occupancy learning with spatiotemporal memory," in *ICCV*, 2025.

[24] X. Yi, Z. Wu, Q. Shen, Q. Xu, *et al.*, "Mvgamba: Unify 3d content generation as state space sequence modeling," in *NeurIPS*, 2025.

[25] J. Li, X. He, C. Zhou, X. Cheng, Y. Wen, and D. Zhang, "Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers," in *ECCV*, 2024.

[26] M. Berman, A. Rannen 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 *CVPR*, 2018.

[27] Z. Murez, T. Van As, J. Bartolozzi, A. Sinha, *et al.*, "Atlas: End-to-end 3d scene reconstruction from posed images," in *ECCV*, 2020.

[28] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[32] 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:2307.01492*, 2023.

[33] Q. Shen, Z. Wu, X. Yi, P. Zhou, H. Zhang, S. Yan, and X. Wang, "Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction," *arXiv preprint arXiv:2403.18795*, 2024.

[34] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *ICML*, 2024.