RoGaussianOCC: An Accurate and Robust Gaussian Occupancy Prediction in Autonomous Driving

Gongjin Lan✉, *Senior Member, IEEE*, Jianfa Ma, Ziyue Zhao, Qining Qi, Yechong Liu and Qi Hao

***Abstract*—In autonomous driving, occupancy prediction is an emerging and mainstream perception method that predicts the spatial occupancy and semantics of 3D voxel grids around the autonomous vehicle from image inputs, providing high robustness in perception, particularly in adverse driving scenes. However, occupancy prediction generally requires dense occupancy representations with redundant computing. Recently, Gaussian-based occupancy prediction has shown state-of-the-art performance in autonomous driving. However, the existing Gaussian occupancy predictions remain inefficient and non-robust (e.g., occlusion scenarios), with several issues including inefficient Gaussian representation, distortion of the Gaussian distribution, and sparse supervision signals. In this work, we propose a novel Gaussian occupancy prediction, called RoGaussianOCC, to solve these issues and improve efficiency and robustness by considering the three key novel modules: 1) Gaussian fusion, 2) Gaussian feature aggregation, and 3) perspective supervision. We conducted an ablation study and compared RoGaussianOCC with state-of-the-art occupancy predictions on the well-known nuScenes dataset. The results show that the proposed three novel modules contribute to the improvement of occupancy prediction in terms of accuracy and robustness, and our RoGaussianOCC outperforms the existing occupancy predictions (e.g., GaussianFormer). Finally, we release the open-source code, video description, and datasets to facilitate the development of Gaussian occupancy prediction in autonomous driving.**

***Index Terms*—Autonomous driving, Occupancy prediction, Gaussian splitting, Gaussian fusion.**

I. Introduction

Autonomous driving generally refers to the technology system that controls the vehicle to reach the destination by recognizing the external environment without driver intervention [1]. Current autonomous driving systems significantly benefit from data-driven deep neural networks for powering critical functions such as perception, decision-making, and planning. Perception is the foundational component of current autonomous driving systems (including end-to-end methods) to understand driving scenes, on which path planning and decision-making rely entirely on the accuracy and reliability of real-time environmental understanding. Bird's-eye view (BEV) perception is the mainstream perception pattern, offering advantages in spatial understanding, sensor fusion, decision-making, and system robustness due to its absolute scale and lack of occlusion in describing environments [2]. Particularly, the general BEV-based perception methods often lack robustness in adverse driving scenes typically including the scenes of 1) sudden illumination changes (e.g., oncoming headlight glare, low-angle sun glare, black and white hole effect of tunnel portals), 2) adverse weather conditions (e.g., heavy rain, dense fog, heavy snow, and dust storm), as shown in Figure 1. Fundamentally, the perception challenges of adverse driving scenes are rooted in the limitations of sensor dynamic range and the constraints of data generalization boundaries, commonly referred to as out-of-distribution (OOD) issues. Consequently, the general BEV-based methods struggle to correctly identify traffic objects in these typically adverse driving scenes, thereby resulting in traffic accidents involving autonomous vehicles.

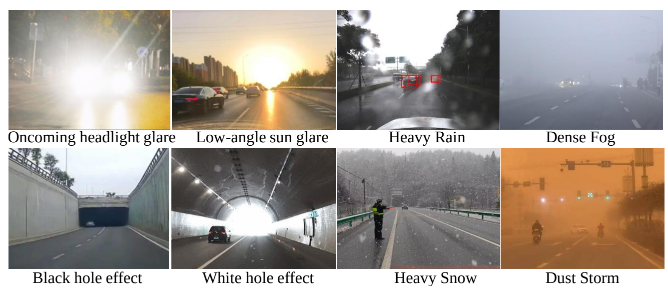

Fig. 1: The adverse scenes of autonomous driving, including the scenes of 1) sudden illumination changes (e.g., oncoming headlight glare, low-angle sun glare, black hole effect (at the entrance), and white hole effect of tunnel portals), 2) adverse weather scenes (e.g., heavy rain, dense fog, heavy snow, and dust storm). (Source: Public datasets and the internet)

Recently, occupancy perception has emerged as a mainstream perception system for autonomous driving, which observes and understands the spatial 3D occupancy and semantics of 3D voxel grids surrounding the autonomous vehicle from visual inputs, where each voxel is classified as occupied or free [3], [4]. Occupancy prediction can represent arbitrarily shaped traffic objects (such as a fallen tree or a leaking truck) that don't fit into standard 3D bounding boxes. It provides the most critical geometric collision risk information, ensuring the vehicle avoids physical collisions and provides high robustness perception, particularly in adverse driving scenes, which are viewed as the next evolution beyond 3D object detection. However, the existing occupancy perception methods generally have several issues, including dense scene representations and

This work was supported in part by the National Natural Science Foundation of China (62261160654), the Shenzhen Fundamental Research Program (JCYJ20220818103006012, KJZD20231023092600001). Corresponding author: Gongjin Lan (e-mail: langj@sustech.edu.cn).

Gongjin Lan, Jianfa Ma, Ziyue Zhao, Qining Qi, Yechong Liu, and Qi Hao are with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China.

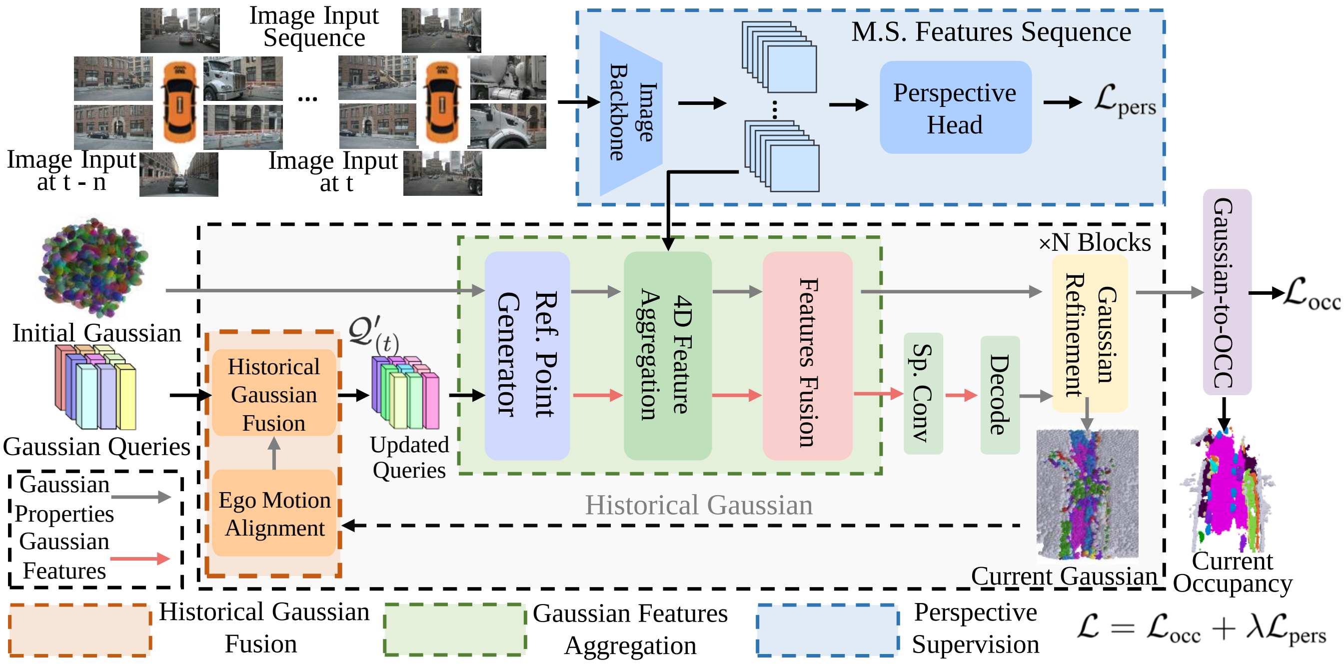

Fig. 2: The framework of our Gaussian occupancy prediction, RoGaussianOCC, contains three main parts: 1) historical Gaussian fusion, 2) Gaussian feature aggregation, and 3) perspective supervision.

redundant computation.

Recently, 3D Gaussians have been applied to occupancy prediction to solve these issues, demonstrating state-of-the-art performance in autonomous driving [5], [6], [7], [8]. In our preliminary experiments, we observed the performance of the well-known Gaussian occupancy predictions (e.g., GaussianFormer [5]) and analyzed the key issues in these methods. Our preliminary experimental results show that existing Gaussian occupancy prediction remains inefficient and non-robust (e.g., occlusion scenarios) with several issues, including 1) inefficient Gaussian representation, 2) Gaussian distribution distortion, and 3) sparse supervision. In this work, we investigate the three key issues in the current Gaussian occupancy prediction for autonomous driving.

1) Inefficient Gaussian representation. The scale of the voxel spatial position in local-aware sparse occupancy prediction is limited to a specific area. Specifically, the spatial updating scale limits the initialized Gaussians in empty regions.
2) Gaussian distribution distortion. The existing Gaussian occupancy prediction primarily updates the Gaussian parameters using the current frame, which is insufficient for accurately predicting the Gaussian distribution of occluded areas and the 3D structure of partially occluded objects.
3) Sparse supervision. The current Gaussian occupancy prediction samples image features sparsely, with only a few areas of interest being supervised, resulting in a sparse supervision signal for the training backbone.

In this work, we propose a novel Gaussian occupancy prediction method, called RoGaussianOCC, to solve three key issues that improve efficiency and robustness by considering three key modules: 1) historical Gaussian fusion, 2) Gaussian feature aggregation, and 3) perspective supervision, as shown in Figure 2. To validate the performance, we conduct an ablation study and compare RoGaussianOCC to state-of-the-art occupancy predictions on the well-known nuScenes dataset. The results show that our RoGaussianOCC outperforms existing well-known occupancy predictions, demonstrating that the proposed three novel modules in our RoGaussianOCC contribute to higher accuracy and robustness in occupancy predictions. Finally, the open source code, video description, and datasets are available on the page [1] to facilitate Gaussian occupancy prediction in autonomous driving. The main contributions of this work can be summarized as:

1) A historical Gaussian fusion that merges current Gaussian features with historical Gaussian spatial distribution, introducing prior Gaussian features to update the spatial updating scale and reduce inefficient Gaussians in empty regions, thereby improving Gaussian representation efficiency.
2) A Gaussian feature aggregation to alleviate the Gaussian distribution distortion in occlusion scenarios by fusing temporal information, which introduces a 4D spatiotemporal sampling of continuous image inputs and aggregates Gaussian query sequences to improve occupancy prediction in occlusion scenarios.
3) A novel perspective supervision as an auxiliary loss to improve the supervision density of the areas of interest during training, thus improving occupancy prediction accuracy.

The rest of this paper is organized as follows. In section II, we review the related work of occupancy prediction in autonomous driving. Our method is presented in section III,

[1] https://github.com/Pluviophil3/RoGaussianOCC

including historical Gaussian fusion, Gaussian feature aggregation, and perspective supervision. section IV addresses the datasets, evaluation, and experimental setup. The results are presented in section V. Finally, we provide an in-depth discussion in section VI and a conclusion in section VII.

II. Related Work

We review the related studies from 1) the state-of-the-art occupancy prediction and 2) the Gaussian occupancy prediction in autonomous driving.

Occupancy Prediction Many studies have investigated occupancy prediction for autonomous driving. Early, Tesla announced the camera-only occupancy network to understand the 3D environment according to surround-view RGB images for autonomous driving. Image-based occupancy predictions are favored for practical deployment due to their low computational and hardware costs [4]. Many remarkable occupancy networks were proposed to perceive 3D scenes, such as SurroundOcc [9], Occ3D [10], OccFormer [11], SparseOcc [12]. Specifically, SurroundOcc [9] constructs image features at various levels and utilizes 2D-3D attention to elevate low-resolution, high-level semantic features to a coarse-grained 3D voxel space. Occ3D [10] proposed an incremental token selection strategy to selectively choose foreground and uncertain voxel tokens during cross-attention computation for adaptive and efficient computation without sacrificing accuracy. OccFormer [11] proposed a dual-path Transformer to encode 3D voxels in different directions and utilize a mask classification model for occupancy prediction. SparseOcc [12] introduced a sparse voxel decoder to reconstruct the scene with the sparse geometric structure by modeling only the non-empty regions, significantly reducing the computing. Although these methods perform remarkably well in occupancy prediction, they employ dense voxels as scene representations, which overlook the occupancy sparsity of natural driving scenes and thus result in a significant amount of redundant computing.

Gaussian Occupancy Prediction Recently, 3D Gaussians have been applied to occupancy prediction to improve occupancy representation efficiency and reduce redundant computing. GaussianFormer [5] proposed the 3D Gaussians as an efficient object-centric representation for driving scenes. Furthermore, an efficient Gaussian-to-voxel splatting generates 3D occupancy predictions by aggregating the neighboring Gaussians. However, GaussianFormer does not consider the continuity of driving scenarios and ignores the prior information from historical frames. GaussianFormer-2 [13] proposed a probabilistic Gaussian superposition model, which interprets each Gaussian as a probability distribution of its neighborhood being occupied and conforms to probabilistic multiplication to derive the overall geometry. Although GaussianFormer-2 optimized the computational efficiency of GaussianFormer, it remains limited to single-frame occupancy prediction. GaussianWorld [6], [14] proposed a Gaussian world model to exploit these priors and infer scene evolution in 3D Gaussian space, which reformulates 3D occupancy prediction as a 4D occupancy forecasting problem conditioned on the current sensor input. GSRender [7] employs 3D Gaussian splatting to simplify the sampling process for occupancy prediction. EmbodiedOcc [15], [16] is a Gaussian-based framework that formulates an embodied 3D occupancy prediction task to target this practical scenario, which initializes the global scene with uniform 3D semantic Gaussians and progressively updates local regions observed by the embodied agent. GS-Occ3D [17] optimizes an explicit occupancy representation using an Octree-based Gaussian Surfel formulation for ensuring efficiency and scalability. OccGS [8] developed a cumulative Gaussian-to-3D voxel splatting method for occupancy prediction using semantic and geometric-aware Gaussian splatting. VoxelSplat [18] leverages enhanced semantics supervision through 2D projection and scene flow learning to improve the accuracy of both semantic occupancy and scene flow estimation. Although these studies have investigated the Gaussian occupancy prediction, the existing Gaussian occupancy predictions remain inefficient and non-robust with several issues including inefficient Gaussian representation, distortion of the Gaussian distribution, and sparse supervision signal. Specifically, they have not considered the global temporal modeling for driving scenes, particularly occlusion scenes.

III. Methodology

To address the accurate and robust Gaussian occupancy prediction, we address the details of our RoGaussianOCC from the three key modules: 1) historical Gaussian fusion, 2) Gaussian feature aggregation, and 3) perspective supervision.

A. Historical Gaussian Fusion

Unlike the traditional BEV (2D) or voxel-based (3D) perception models, Gaussian occupancy prediction generates a Gaussian distribution of objects to update the Gaussian distribution of the driving scene by modeling the offset of each Gaussian ellipsoid across the frames [19], [20]. Gaussian representations account for the complexity of different regions, considering the uneven distribution of objects in 3D scenes. For example, the empty regions do not require an excessive number of Gaussians, while the regions with dense objects require an adequate density of Gaussians. The existing Gaussian occupancy predictions, such as GaussianFormer [5], typically employ sparse convolution and deformable attention to update the offset of the Gaussian mean within the attention region of the image [20]. However, the receptive field in these methods is constrained to the regions of interest within the image, and the updating scale of Gaussians is thus limited, resulting in low Gaussian representation efficiency. In this work, we propose a novel scheme that leverages historical Gaussian to refine Gaussian representations, which consists of 1) ego-motion alignment and 2) historical Gaussian fusion, as shown in Figure 3. Specifically, our method fuses historical Gaussian queries with initial Gaussian queries, utilizing the prior historical Gaussian properties with temporal information to update current Gaussians.

First, we extract the Gaussian properties from the previous frame ($\mathcal{G}_{(t-1)}$) and apply an alignment function (f_a) to align the mean and rotation of historical Gaussians of the current

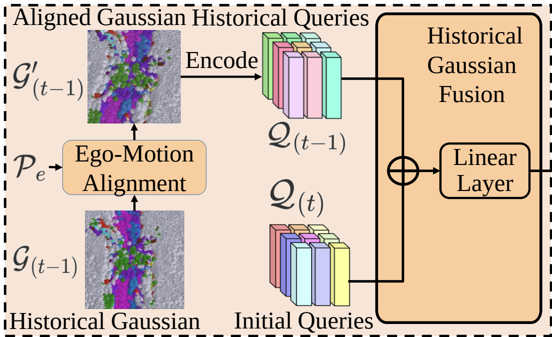

Fig. 3: The framework of historical Gaussian fusion consists of two parts: 1) ego-motion alignment and 2) historical Gaussian fusion.

frame. Specifically, we convert each Gaussian $g \in \mathcal{G}_{(t-1)}$ to the aligned Gaussian features g' by the alignment function (f_a):

$$g' = f_a(g_{(t-1)}) = m_{(t-1)} + \Delta m, s_{(t-1)}, \Delta r \otimes r_{(t-1)}, c_{(t-1)} \quad (1)$$

where $m_{(t-1)}, s_{(t-1)}, r_{(t-1)}, c_{(t-1)}$ denote the mean, scale, rotation, and semantic label of the Gaussian $g_{(t-1)}$; Δm, Δr denote the offsets of Gaussian mean and rotation quaternion caused by movement, which are calculated by the extrinsic parameters ($\mathcal{P}_e$), and $\otimes$ represents multiplication of quaternion. Thus, the previous Gaussian properties are correctly aligned with the current frame. Finally, we encode the aligned Gaussian $\mathcal{G}'_{(t-1)}$ as the historical query $\mathcal{Q}_{(t-1)}$.

Second, we design a historical Gaussian fusion that integrates the aligned historical Gaussian query $\mathcal{Q}_{(t-1)}$ with the current Gaussian query $\mathcal{Q}_{(t)}$ to a novel Gaussian query $\mathcal{Q}'_{(t)}$ with the enriched historical Gaussian.

$$\mathcal{Q}'_{(t)} = \mathrm{F}_{map}(\mathcal{Q}_{(t)} \oplus \mathcal{Q}_{(t-1)}) \quad (2)$$

where F_{map} and $\oplus$ denote the mapping function by the feed-forward networks and concatenation, respectively. Gaussians with prior information, after historical Gaussian fusion, present previous scenes, which facilitates the optimized distribution of Gaussians to represent complex objects, thereby improving the efficiency of Gaussian representation. The fused Gaussian representation maintains temporal and spatial continuity, as the Gaussian properties are propagated and continuously integrated over both temporal and spatial dimensions, thereby enhancing the stability and consistency of the Gaussian representation in dynamic continuous scenes [21].

B. Gaussian Feature Aggregation

The single-frame input in the existing Gaussian occupancy prediction is inherently limited in perspective and often fails to provide accurate predictions for occluded regions. In autonomous driving scenarios, occlusions frequently occur due to the complexity of the environment, but previously visible regions may reappear in earlier frames as the vehicle moves. Temporal feature aggregation across multiple frames can reduce prediction errors caused by occlusions. To address this issue, we propose Gaussian feature aggregation, incorporating historical and current image inputs, alleviating occupancy prediction failure for occluded scenarios and enhancing temporal scene perception.

Specifically, we design a Gaussian feature aggregation by referring to the basic framework of GaussianFormer [5]. Our Gaussian feature aggregation consists of three parts: 1) reference points generator, 2) 4D feature aggregation, and 3) feature fusion, as shown in Figure 4.

a) Reference points generator: The generator produces reference points based on the Gaussian features and query features of the current frame. Specifically, it integrates each Gaussian $g_i \in G$ with its corresponding query vector $q_i \in \mathcal{Q}'_{(t)}$, followed by a linear layer to generate 3D reference points (r_i).

$$r_i = G(q_i, g_i), r_i \in \mathcal{R} \quad (3)$$

where G represents the operation of reference points generator by linear layer, and $\mathcal{R}$ denotes the set of all generated reference points of the Gaussians $\mathcal{G}$.

b) 4D feature aggregation: We propose a 4D feature aggregation to enable multi-view and multi-scale feature extraction across temporal and spatial dimensions. It is necessary to adjust the reference points to positions that align with the temporal sequence frames, considering the different vehicle positions across frames in the multi-scale feature sequence. We align the generated reference points $\mathcal{R}$ by considering the spatial offset with the camera extrinsic parameters $\mathcal{P}_e$.

$$\mathcal{R}' = Align(\mathcal{R}, \mathcal{P}_e) \quad (4)$$

The aligned reference points $\mathcal{R}'$ are projected to feature maps according to the camera's intrinsic parameters $\mathcal{P}_i$ and extrinsic parameters $\mathcal{P}_e$. Finally, the Gaussian feature (f_t) can be calculated with the aligned reference points $\mathcal{R}'$ and feature map $\mathcal{F}$ by the deformable attention.

$$f_t(\mathcal{R}', \mathcal{Q}, \mathcal{F}; \mathcal{P}_e, \mathcal{P}_i) = \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \sum_{n=1}^{\mathcal{N}} \mathrm{DA}(\pi(\mathcal{R}'_n; \mathcal{P}_i, \mathcal{P}_e), \mathcal{F}_m) \quad (5)$$

where $\mathcal{N}$, $\mathcal{M}$, $\pi(\cdot)$, $\mathcal{F}_m$, DA denote the number of Gaussians, the number of feature maps, the transformation function from world to pixel coordinates on feature maps, the m-th feature map, the deformable attention function, respectively.

c) Features fusion: To enhance the robustness of the occupancy prediction in occluded scenes, we design the feature fusion to extract the final Gaussian feature (f) of the current frame by integrating the historical Gaussian feature sequence $f_{t-l}, \cdots, f_{t-1}, f_t$.

$$f(\mathcal{R}, \mathcal{Q}, \mathcal{F}; \mathcal{P}_e, \mathcal{P}_i) = \mathcal{O}\left(f_{t-l} \oplus \cdots \oplus f_{t-1} \oplus f_t\right) \quad (6)$$

where l denotes the sequence length, $\mathcal{O}$ is the linear layer operation.

C. Perspective Supervision

In Gaussian occupancy prediction, the sparsity of reference points in deformable attention hinders the effective backpropagation of gradients to the backbone. The current Gaussian occupancy prediction samples image features sparsely, with only a few areas of interest being supervised, resulting in sparse supervision signals for the training backbone. These

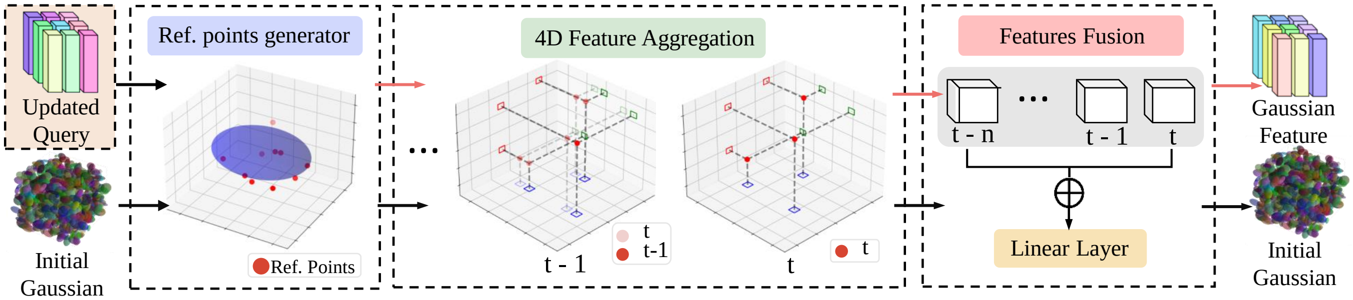

Fig. 4: Gaussian feature aggregation: 1) reference points generator, 2) 4D feature aggregation, and 3) features fusion.

limitations result in insufficient convergence capability during the early training stages and restrict feature extraction in Gaussian occupancy prediction [22].

We propose perspective loss as an auxiliary supervision signal, directly guiding the feature extraction process in the backbone to enhance the supervision density of areas of interest during training, as shown in Figure 2. The perspective supervision head estimates the depth for each pixel in the feature maps and aligns it with the depth information derived from the LiDAR point cloud in the dataset. The perspective loss ($\mathcal{L}_{pers}$) can be calculated by:

$$\mathcal{L}_{\text{pers}} = \frac{1}{\mathcal{S}} \sum_{i=1}^{\mathcal{S}} \left| d_i - d_i^{\text{gt}} \right| \tag{7}$$

where $\mathcal{S}$, d, and d^{gt} denote the total number of pixels, the predicted depth, and the ground truth of depth, respectively. Finally, we formulate the overall loss function by considering both occupancy and perspective loss:

$$\mathcal{L} = \mathcal{L}_{\text{occ}} + \lambda \mathcal{L}_{\text{pers}} \tag{8}$$

where $\mathcal{L}_{\text{occ}}$ and $\mathcal{L}_{\text{pers}}$ represent the occupancy loss and the perspective loss, respectively, and λ is a hyperparameter that balances the influence of the occupancy loss and perspective loss.

IV. Experiments

We conducted a series of experiments on the well-known nuScenes dataset for autonomous driving to validate the accuracy and robustness of RoGaussianOCC. In this section, we address the datasets, measurement metrics, and experimental setup.

A. Datasets

1) nuScenes: We conduct experiments on the well-known large-scale autonomous driving dataset nuScenes [23], which contains 1000 scenes of 20 seconds each, and 1400000 images. There are 28130 samples for training, 6019 samples for validation, and 6008 samples for testing. The 3D instances are annotated into 17 non-empty classes, including cars, trucks, buses, trailers, pedestrians, motorcycles, and bikes. In this work, 1000 scenes are divided into 850 scenes for training and 150 scenes for testing. The dense annotations provided by SurroundOcc [9] are used as ground truth for calculating occupancy loss. The perspective loss is calculated with the depth generated by the LiDAR point cloud from the nuScenes dataset during the training stage.

2) Occlusion Dataset: To evaluate the occupancy prediction of RoGaussianOCC for occlusion scenes, we manually selected 227 samples with occlusion challenges (e.g., poor visibility, multiple vehicles) from the nuScenes validation set as an occlusion dataset and conducted an ablation study on the occlusion dataset to validate the performance of Gaussian feature aggregation in our RoGaussianOCC for occlusion scenes.

B. Measurement Metrics

In this work, we use the Intersection over Union (IoU) and mean IoU (mIoU) to evaluate the performance of RoGaussianOCC.

$$\text{mIoU} = \frac{1}{|C'|} \sum_{i \in C'} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{9}$$

$$\text{IoU} = \frac{TP_i}{TP_{\neq C_e} + FP_{\neq C_e} + FN_{\neq C_e}} \tag{10}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. C' denotes all non-empty semantic labels, while C_e represents the empty label.

Furthermore, we use the percentage of Gaussians in correct positions ($\mathcal{P}_p$) and semantic [20] ($\mathcal{P}_s$) to evaluate the positional accuracy and semantic prediction accuracy of Gaussians, respectively, which can be calculated by:

$$\mathcal{P}_p = \frac{\mathcal{N}_{in}}{\mathcal{N}}, \quad \mathcal{P}_s = \frac{\mathcal{N}_{TP}}{\mathcal{N}_{in}} \tag{11}$$

where $\mathcal{N}$ represents the total number of Gaussians, $\mathcal{N}_{in}$ denotes the number of Gaussians whose centers reside within occupied (non-empty) voxels, and $\mathcal{N}_{TP}$ indicates the number of Gaussians whose centers are located within occupied voxels and have correctly predicted semantics (i.e., Ture Positive).

C. Experimental Setup

We use an input image resolution of 900×1600 and the initialized ResNet101-DCN [24] with pre-trained weights from FCOS3D [25] as the backbone, ensuring consistency with well-known studies such as GaussianFormer [5] and GaussianWorld [6]. We conducted hyperparameter tuning experiments for $\lambda = 0.1, 0.2, 0.3, 0.5, 1$ and finally selected $\lambda = 0.3$ for a best IoU and mIoU. Furthermore, we use the AdamW optimizer (a weight decay of 0.1) to optimize the learning rates during training with a maximum learning rate of $4e^{-4}$ and a batch size of 1 [26]. With these parameters, our RoGaussianOCC was trained for 20 epochs on the GPU system with 6 NVIDIA V100S, which took approximately 110 hours.

V. Results

Here, we present the experimental results of the ablation study, comparisons of our RoGaussianOCC with various occupancy prediction models, and visualization of accuracy and robustness for various adverse scenes.

A. Ablation Study

We conduct an ablation study to provide a comprehensive analysis of the three modules in RoGaussianOCC with a Gaussian number ($\mathcal{N}$=25600) on the nuScenes dataset. Specifically, we ablate historical Gaussian fusion, Gaussian feature aggregation, and perspective supervision to evaluate their contributions to the proposed RoGaussianOCC. The results in Table I demonstrate that each module contributes positively to the accuracy of occupancy prediction. Specifically, disabling the historical Gaussian fusion module resulted in a performance drop of 1.18% in IoU and 1.88% in mIoU. Disabling the Gaussian feature aggregation module resulted in a decrease of 1.09% in IoU and 1.73% in mIoU. Furthermore, the disabling of perspective supervision caused a significant reduction of 2.46% IoU and 5.74% mIoU. Importantly, perspective supervision contributes the most significant improvement as it substantially improves feature extraction in the backbone during training. Finally, our RoGaussianOCC outperforms the baseline (GaussianFormer) by 9.02% in IoU and 26.4% in mIoU.

Method	IoU(%)	mIoU(%)
GaussianFormer	28.72	16.00
Ours (HGF+GA+PS)	**31.31(9.02%↑)**	**20.22(26.4%↑)**
Ours (GA+PS)	30.94(1.18%↓)	19.84(1.88%↓)
Ours (HGF+PS)	30.97(1.09%↓)	19.87(1.73%↓)
Ours (HGF+GA)	30.54(2.46%↓)	19.06(5.74%↓)

TABLE I: The results of ablation studies on three modules of our RoGaussianOCC (Gaussians number ($\mathcal{N}$=25600)).

Furthermore, we conducted ablation experiments to analyze the contributions of historical Gaussian fusion and Gaussian feature aggregation on the efficiency of Gaussian representation by observing the positional and semantic accuracy of Gaussians (i.e., $\mathcal{P}_p$ and $\mathcal{P}_s$). The results in Table II show that both historical Gaussian fusion and Gaussian feature aggregation enhance the positional and semantic Gaussian accuracy, which thus improves the efficiency of Gaussian representation and better occupancy prediction accuracy.

Method	$\mathcal{P}_p$(%)	$\mathcal{P}_s$(%)
Ours (HGF+PS)	12.05 (11.9%↓)	74.66 (2.70%↓)
Ours (GA+PS)	13.18 (3.65%↓)	76.54 (0.25%↓)
Ours (HGF+GA+PS)	**13.68**	**76.73**

TABLE II: The contributions of historical Gaussian fusion and Gaussian feature aggregation to positional and semantic accuracy of Gaussians.

The Gaussian number ($\mathcal{N}$) is a key parameter that directly affects accuracy and the computation of Gaussian occupancy prediction. In general, a larger number of Gaussians contributes to higher accuracy and increased computational complexity. We test various Gaussian numbers and observe their accuracy (IoU and mIoU), as shown in Table III. The results show that our RoGaussianOCC with a Gaussian number ($\mathcal{N}$=25600) performs a significant improvement of 9.02% IoU and 26.4% mIoU compared to the well-known GaussianFormer. Furthermore, our RoGaussianOCC with a larger Gaussian number ($\mathcal{N}$=51200) performs a 12.1% IoU and 29.1% mIoU improvement, but with a 13.6% memory improvement. Although the larger Gaussian number leads to higher accuracy, it requires larger computing. We analyze the various Gaussian numbers and their accuracy by the Pareto front, as shown in Figure 5. The Pareto front results show that our RoGaussianOCC, with Gaussian numbers of 25600 and 51200, dominates other Gaussian occupancy predictions. To balance the accuracy and computation, we choose the Gaussian number as 25600 in RoGaussianOCC.

Method	$\mathcal{N}$	Memory	IoU(%)	mIoU(%)
Gaussian Former	25600	4850MB	28.72	16.00
	144000	6229MB(28.4%↑)	29.83(3.86%↑)	19.10(19.3%↑)
Gaussian Former-2*	25600	3063MB (37.8%↓)	30.56(6.41%↑)	20.02(25.1%↑)
Ours	6400	4652MB(4.08%↓)	28.50(0.77%↓)	18.52(15.8%↑)
	25600	4680MB(3.51%↓)	31.31(9.02%↑)	20.22(26.4%↑)
	51200	5508MB(13.6%↑)	32.20(12.1%↑)	20.65(29.1%↑)

TABLE III: The results of various Gaussian numbers ($\mathcal{N}$) in the well-known Gaussian occupancy predictions and our RoGaussianOCC.

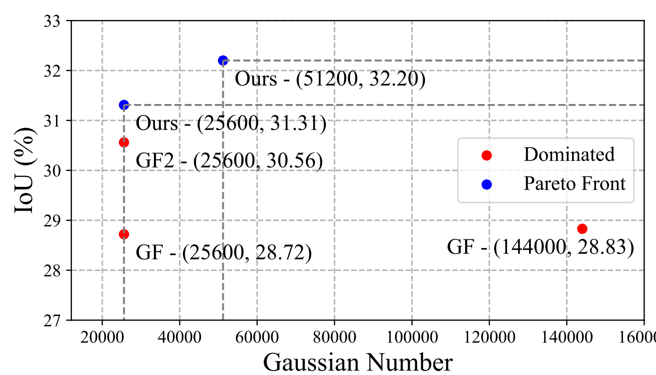

Fig. 5: The Pareto front results of various Gaussian numbers and accuracy. Ours (blue) dominates others (red).

Finally, we manually selected 227 samples as an occlusion dataset and conducted an ablation study on the occlusion dataset to further validate the performance of Gaussian feature aggregation in our RoGaussianOCC for occlusion scenes. The results in Table V show that the historical Gaussian fusion leads to a 9.99% IoU and 14.2% mIoU improvements. The Gaussian feature aggregation contributes to improvements of 1.17% and 2.23% in IoU and mIoU, respectively.

Method	Memory (MB)	IoU(%)	mIoU(%)	barrier	bicycle	bus	car	motor.	pedes.	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [28]	-	23.96	7.31	4.03	0.35	8.00	8.04	0.28	1.16	0.67	4.01	4.35	27.72	5.20	15.13	11.29	9.03	14.86
Atlas [29]	-	28.66	15.00	10.64	5.68	19.66	24.94	8.84	6.47	3.28	10.42	16.21	34.86	15.46	21.89	20.95	11.21	20.54
BEVFormer [30]	25100	30.50	16.75	14.22	6.58	23.46	28.28	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer [31]	29000	30.86	17.10	15.96	5.31	23.86	27.32	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
OccFormer [11]	>22400†	31.39	19.03	18.65	10.41	23.92	30.29	14.19	**13.59**	**10.13**	12.49	20.77	38.78	19.79	24.19	22.21	13.48	21.35
GaussianFormer‡ [5]	6229	29.83	19.10	19.52	11.26	26.11	29.78	13.83	12.58	8.67	12.74	21.57	39.63	23.28	24.46	22.99	9.59	19.12
GaussianFormer-2*[20]	3063	30.56	20.02	20.15	12.99	**27.61**	30.23	15.31	12.64	9.63	13.31	**22.26**	39.68	23.47	25.62	23.20	12.25	20.73
Ours(25600)	4680	31.31 (4.96%↑)	20.22 (5.86%↑)	19.38	13.05	27.01	30.16	15.52	13.11	9.29	**13.83**	21.95	39.92	**23.88**	25.72	23.63	13.33	22.71
Ours(51200)	5508	**32.20(7.95%↑)**	**20.65(8.12%↑)**	**20.28**	**13.05**	27.30	**30.48**	**16.21**	13.47	9.71	13.42	22.25	**40.39**	23.81	**26.26**	**24.16**	**14.85**	**23.80**

TABLE IV: The comparison between our RoGaussianOCC and the state-of-the-art vision-based occupancy predictions on nuScenes datasets, where the performance of GaussianFormer is the baseline. Memory Data is reported from GaussianFormer-2[20]. † is referenced in OctreeOCC [27]. ‡ and * represent the Gaussian numbers of 144000 and 25600, respectively.

Method (*+PS)	IoU(%)	mIoU(%)	Barrier	Car	Ped.	Tra.
Ours (HGF)	27.04 (9.99%↓)	18.86 (14.2%↓)	15.96	26.48	10.83	8.90
Ours (GA)	29.69 (1.17%↓)	21.49 (2.23%↓)	19.63	30.29	12.81	12.58
Ours (HGF+GA)	**30.04**	**21.98**	**21.55**	**30.52**	**13.36**	**13.08**

TABLE V: The results of RoGaussianOCC on the occlusion dataset with common occluded objects.

B. Occupancy Prediction

We compare RoGaussianOCC to the state-of-the-art vision-based occupancy predictions on the nuScenes dataset, as shown in Table IV, where the performance of GaussianFormer is selected as the baseline. The results show that our RoGaussianOCC significantly outperforms the well-known BEV-based methods, such as BEVFormer [30] and TPVFormer [31]. For example, our RoGaussianOCC ($\mathcal{N}$=25600) performs an improvement of 1.46% IoU and 18.25% mIoU compared to TPVFormer. Furthermore, RoGaussianOCC performs with an accuracy of approximately the same or higher than that of occupancy prediction methods with dense grid representation. For example, RoGaussianOCC ($\mathcal{N}$=25600) and RoGaussianOCC ($\mathcal{N}$=51200) perform 6.25% and 8.51% mIoU improvements, respectively, compared to the dense grid representation method OccFormer [11].

Finally, our RoGaussianOCC with the same Gaussian numbers (25600) performs higher accuracy than the existing Gaussian occupancy predictions, such as GaussianFormer [5] and GaussianFormer-2 [20]. Specifically, our RoGaussianOCC

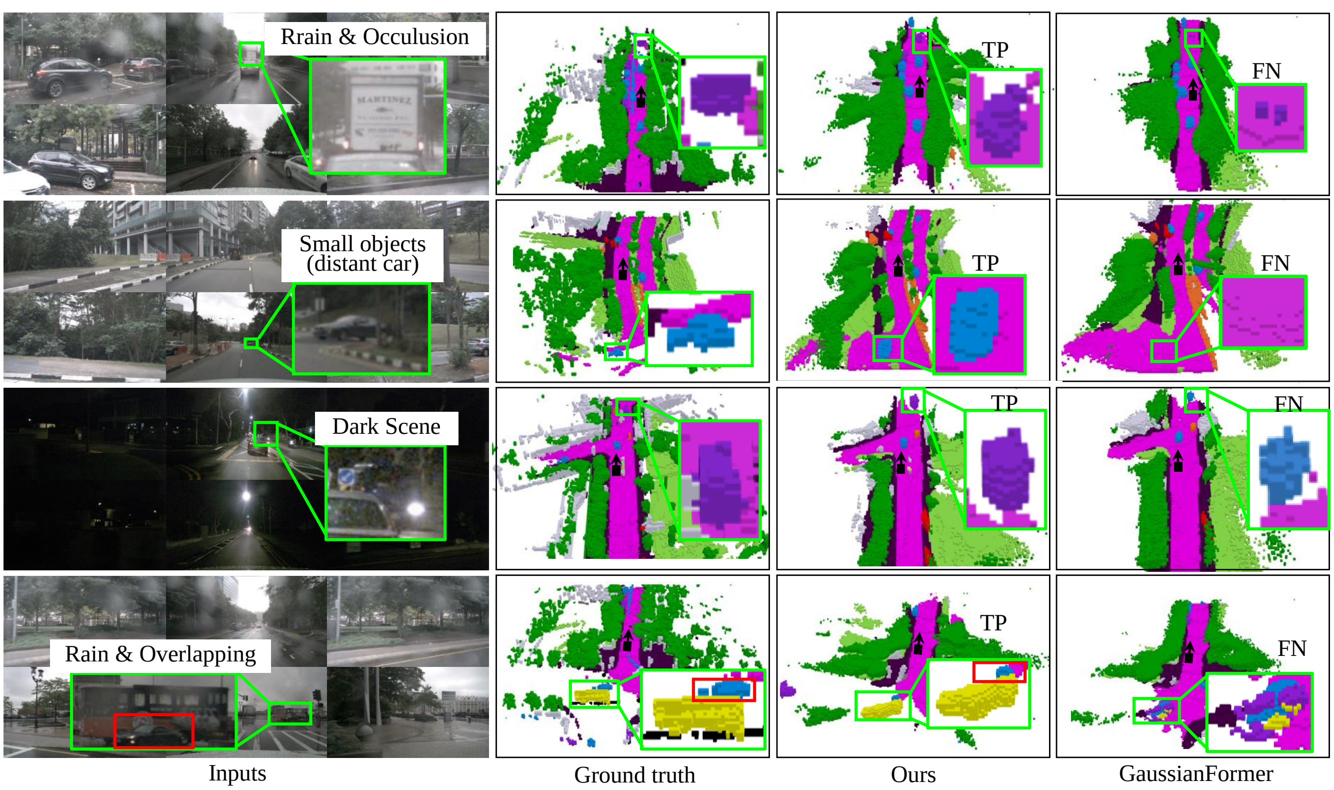

Fig. 6: Visualization of Gaussian representation and occupancy prediction of our RoGaussianOCC on the nuScenes dataset.

with a Gaussian number of 25600 outperforms GaussianFormer with an improvement of 4.96% IoU and 5.86% mIoU, and outperforms GaussianFormer-2 with an improvement of 2.45% IoU and 1.00% mIoU. For a larger Gaussian number of 51200, our RoGaussianOCC achieves an improvement of 7.95% in IoU and 8.12% in mIoU compared to GaussianFormer, respectively.

C. Visualization Results

We conduct a qualitative analysis of the occupancy prediction results of RoGaussianOCC. First, we present the visualization results of occupancy prediction of three representative scenes from nuScenes, as shown in Figure 6. The visual results show that our RoGaussianOCC accurately predicts the occupancy and semantics of small objects and occluded objects. Specifically, the truck in the front view of the first scene (first row in Figure 6) is partially occluded by a car, which is correctly predicted by our RoGaussianOCC but incorrectly by GaussianFormer, as GaussianFormer with single-frame input predicts occupancy without temporal information. A distant car in the second scene (second row in Figure 6) is ignored by GaussianFormer since the distant car is viewed as a small object, while it is predicted by our RoGaussianOCC. The third scene (third row) demonstrates that our RoGaussianOCC accurately predicts occupancy in the dark scene, whereas GaussianFormer predicts occupancy incorrectly. Our GaussianFormer predicts the occupancy for the scene of a car overlapped with a bus (fourth row), which GaussianFormer predicts incorrectly.

Furthermore, we observe the visual results of the Gaussian distribution caused by historical Gaussian fusion, as shown in Figure 7 where the red dots represent the accurate Gaussian whose mean value is in the voxel. The percentage is the ratio of the accurate Gaussians to all Gaussians. Figure 7 shows three various scenes with low to high accurate Gaussians. The visual results show that Gaussians in our RoGaussianOCC are concentrated in the regions of interest that require intensive Gaussian representations, which reduces Gaussian redundancy in empty regions and has higher $\mathcal{P}_p$ scores than GaussianFormer. For example, the Gaussians are concentrated on the road and the objects on both sides that are key occupancies to autonomous driving.

Finally, we present the visual results of Gaussian feature aggregation of three representative scenes in Figure 8. The visual results show that our RoGaussianOCC with Gaussian feature aggregation performs accurate occupancy predictions for occluded objects. In the first scene (first row), the targeted car behind the ego-vehicle is partially occluded, which is correctly predicted by our RoGaussianOCC with GFA and incorrectly predicted without GFA, respectively. In the second

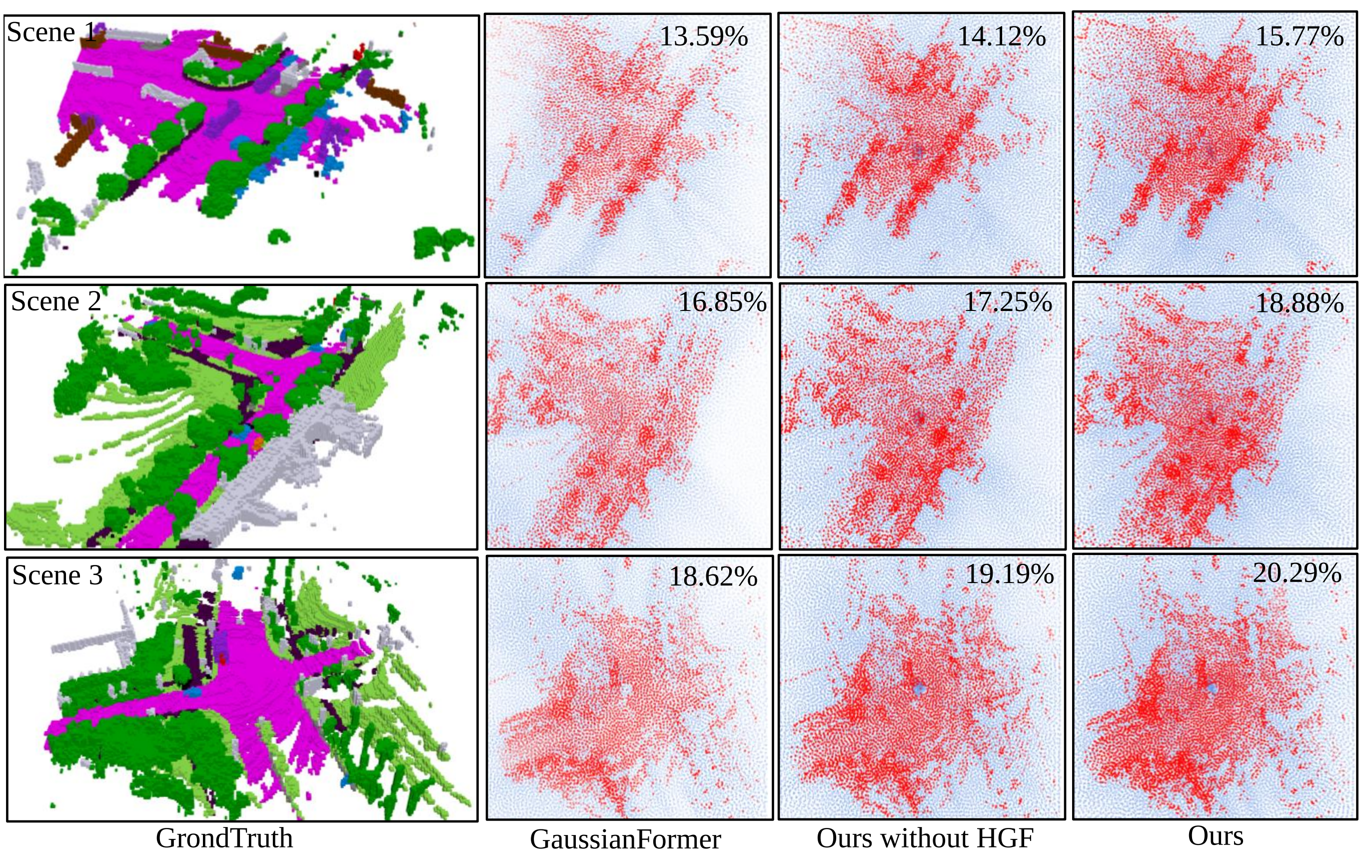

Fig. 7: Visualization of a Gaussian distribution. Our RoGaussianOCC with historical Gaussian fusion reduces redundant Gaussian representations in empty regions and increases the Gaussian density in regions of interest (e.g., roadways). The percentage is the ratio of the accurate Gaussians to all Gaussians. The red and light blue dots represent the accurate Gaussian whose mean value is within the voxel and out of the voxel, respectively.

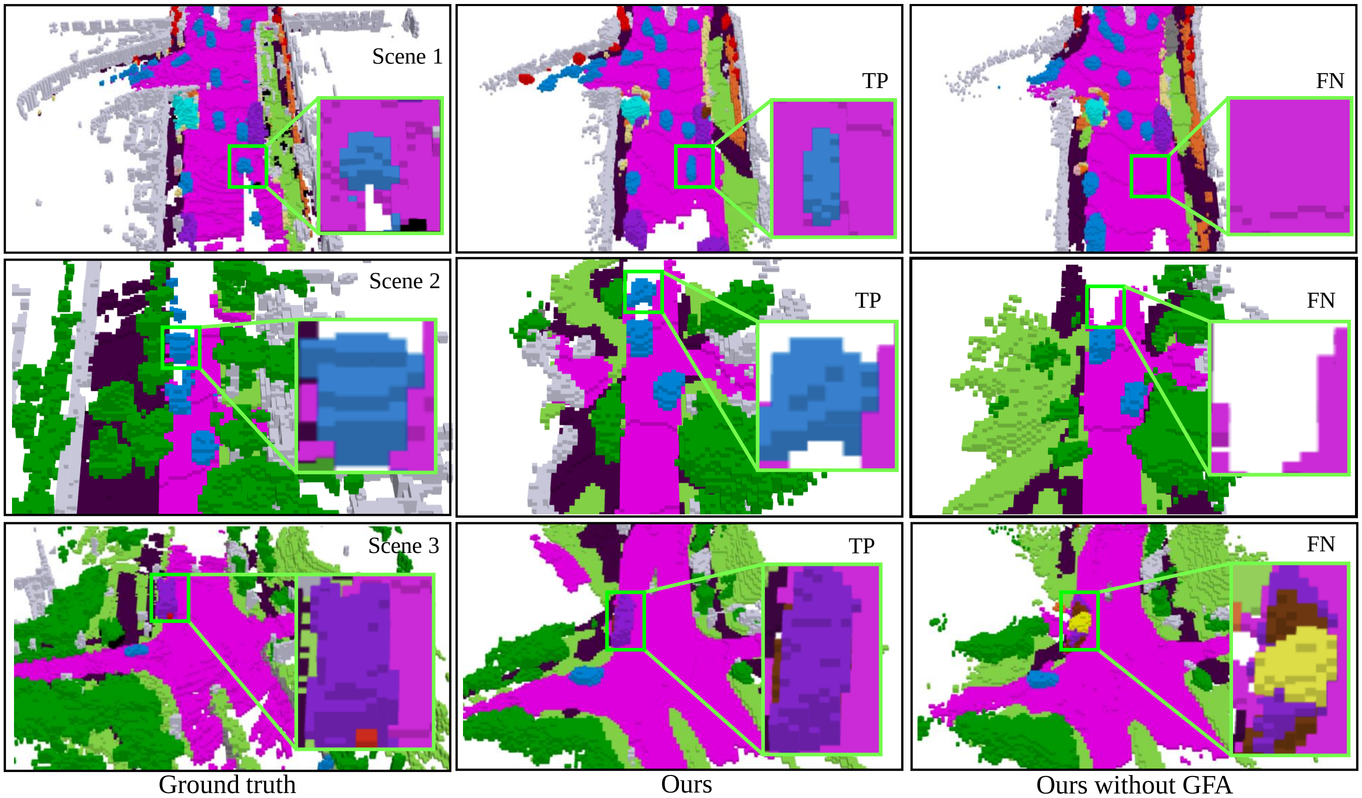

Fig. 8: Visualization of occupancy prediction with Gaussian feature aggregation. TP: true positives, FN: false negatives.

scene, our RoGaussianOCC with GFA accurately predicts occupancy for the challenging scenario of multiple vehicles on a narrow road. In the third scene, our RoGaussianOCC with GFA module correctly predicts the occupancy of a truck at the crossroads.

VI. Discussion

Although our RoGaussianOCC shows remarkable performance in occupancy predictions, there are still many interesting issues that need to be further investigated. First, we observed that Gaussian occupancy prediction methods based on deformable attention, including our RoGaussianOCC, still require further improvement for small objects. This issue is generally caused by two reasons: 1) small objects are assigned fewer Gaussians, resulting in weak feature representation; 2) small objects are often partly occluded and thus are hardly predicted. Second, the Gaussian feature aggregation in our RoGaussianOCC considers the previous keyframes as a form of temporal fusion. However, frames with efficient historical information can be selected based on the vehicle's historical path. In the future, we will select historical keyframes by considering the vehicle's historical path. Third, the ego-motion alignment in the historical Gaussian fusion of our RoGaussianOCC only considers the motion of the ego vehicle, ignoring the motion of other moving objects. Considering that all motion objects may improve Gaussian occupancy predictions.

In the GFA module, the ego-motion alignment only considers the movement of the ego-vehicle but ignores the movement of other traffic objects, which may lead to underfitting results. The movement of other traffic objects may contribute to improving occupancy prediction. Additionally, we employ a fixed temporal sampling frame rate, which results in inefficient Gaussian feature aggregation when the vehicle is moving slowly or stopped. To address this issue, we will utilize the ego-vehicle motion information to model temporal information for key frame sampling, thereby adapting to low- or high-speed driving scenes [32]. Although we introduced perspective supervision to train the backbone, it is directly applied to occupancy prediction and cannot provide supervision for Gaussians. In the future, we will introduce the Gaussian L1 distance and the overlay to comprehensively evaluate the performance. Additionally, we will utilize the occupancy annotation of nuScenes to generate ground truth for Gaussian training.

Although our RoGaussianOCC outperforms other methods, OccFormer outperforms our RoGaussianOCC in terms of pedestrian and traffic cone, as shown in Table IV. We noticed that deformable attention-based Gaussian occupancy prediction methods (RoGaussianOCC) generally perform weaker on small objects (such as pedestrians and traffic cones) than traditional occupancy prediction methods, such as OccFormer. Small objects are assigned a few Gaussians that are hardly to be predicted as occupancy. In contrast, traditional occupancy prediction methods that use voxels as the representation can accurately decode small objects and predict them as occupied. In the future, we will incorporate confidence into Gaussians to enhance the occupancy prediction of small objects.

VII. Conclusions

In this work, we proposed a novel Gaussian occupancy prediction, RoGaussianOCC, to improve the efficiency and robustness by considering three key modules: 1) historical Gaussian fusion, 2) Gaussian feature aggregation, and 3) perspective supervision. We conducted an ablation study to validate the contributions of the three novel modules to the improvement of occupancy prediction and compared RoGaussianOCC to state-of-the-art occupancy predictions (e.g., Gaussianformer, Gaussianformer-2) on the well-known nuScenes dataset. The results demonstrate that our RoGaussianOCC outperforms existing Gaussian occupancy predictions (e.g., GaussianFormer and GaussianFormer-2) in terms of accuracy and robustness for autonomous driving applications. The efficiency and robustness of occupancy prediction significantly provide the cornerstone for path planning and decision-making of autonomous driving. However, our work focused on the perception stage (occupancy prediction), a crucial step in autonomous driving, without validating the performance improvements it brings to the end-to-end autonomous driving system. In the future, we will apply the proposed RoGaussianOCC to validate its contribution to the autonomous driving system and improve the performance of end-to-end autonomous driving.

References

[1] G. Lan, Y. Peng, Q. Hao, and C. Xu, "SustechGAN: Image generation for object detection in adverse conditions of autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[2] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng *et al.*, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2151–2170, 2023.

[3] 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.

[4] 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.

[5] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Gaussianformer: Scene as gaussians for vision-based 3D semantic occupancy prediction," in *European Conference on Computer Vision*. Springer, 2024, pp. 376–393.

[6] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "Gaussianworld: Gaussian world model for streaming 3D occupancy prediction," *arXiv preprint arXiv:2412.10373*, 2024.

[7] Q. Sun, C. Shu, S. Zhou, Z. Yu, Y. Chen, D. Yang, and Y. Chun, "Gsrender: Deduplicated occupancy prediction via weakly supervised 3D gaussian splatting," *arXiv preprint arXiv:2412.14579*, 2024.

[8] X. Zhou, J. Wang, Y. Wang, Y. Wei, N. Dong, and M.-H. Yang, "OccGS: Zero-shot 3D occupancy reconstruction with semantic and geometric-aware gaussian splatting," *arXiv preprint arXiv:2502.04981*, 2025.

[9] 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.

[10] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.

[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] 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.

[13] Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du, and J. Lu, "Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 477–27 486.

[14] T. Feng, W. Wang, and Y. Yang, "Gaussian-based world model: Gaussian priors for voxel-based occupancy prediction and future motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 25 239–25 249.

[15] Y. Wu, W. Zheng, S. Zuo, Y. Huang, J. Zhou, and J. Lu, "EmbodiedOcc: Embodied 3D occupancy prediction for vision-based online scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 26 360–26 370.

[16] H. Wang, X. Wei, X. Zhang, J. Li, C. Bai, Y. Li, M. Lu, W. Zheng, and S. Zhang, "Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 925–934.

[17] B. Ye, M. Qin, S. Zhang, M. Gong, S. Zhu, H. Zhao, and H. Zhao, "Gs-Occ3D: scaling vision-only occupancy reconstruction with gaussian splatting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 25 925–25 937.

[18] Z. Zhu, S. Wang, J. Xie, J.-j. Liu, J. Wang, and J. Yang, "Voxelsplat: Dynamic gaussian splatting as an effective loss for occupancy and flow prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6761–6771.

[19] W. Gan, F. Liu, H. Xu, N. Mo, and N. Yokoya, "GaussianOcc: fully self-supervised and efficient 3D occupancy estimation with gaussian splatting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 28 980–28 990.

[20] Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du, and J. Lu, "Probabilistic gaussian superposition for efficient 3D occupancy prediction," *arXiv preprint arXiv:2412.04384*, 2024.

[21] L. Chambon, E. Zablocki, A. Boulch, M. Chen, and M. Cord, "Gaussrender: Learning 3d occupancy with gaussian rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 27 010–27 020.

[22] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4D v2: recurrent temporal fusion with sparse model," *arXiv preprint arXiv:2305.14018*, 2023.

[23] 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.

[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. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 913–922.

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

[27] Y. Lu, X. Zhu, T. Wang, and Y. Ma, "OctreeOcc: efficient and multi-granularity occupancy prediction using octree queries," 2024.

[28] 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.

[29] Z. Murez, T. Van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3D scene reconstruction from posed images," in *European conference on computer vision*. Springer, 2020, pp. 414–431.

[30] 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 lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[31] 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.

[32] Y. Gao, C. Li, Z. You, J. Liu, Z. Li, P. Chen, Q. Chen, Z. Tang, L. Wang, P. Yang *et al.*, "OpenFly: a versatile toolchain and large-scale benchmark for aerial vision-language navigation," *arXiv e-prints*, pp. arXiv–2502, 2025.

Gongjin Lan (Senior Member, IEEE) is currently a Research Assistant Professor, Associate Researcher at the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. He received his PhD in Artificial Intelligence at VU Amsterdam in 2020, and published over 30 academic papers (>1000 citations, h-index=21) in top journals and conferences, including *IEEE T-IV, IEEE THMS, IEEE TII, IEEE TLT, Swarm and Evolutionary Computation, Knowledge-Based Systems, Scientific Reports, PPSN*, and serves as a active reviewer for 15+ top journals/conferences (including *Nature Communications, IEEE TEVC, IEEE THMS, IEEE TII, IEEE TMC, IEEE RA-L, Swarm and Evolutionary Computation, Scientific Reports, ICRA, IROS, GECCO, IJCNN*), the Guest Editor for a SCI-Q3 journal and the Session Chair at a top IEEE conference, the Secretary and Co-Founder of a IEEE ITSS Chapter. His research interests include autonomous driving, embodied AI, and interdisciplinary AI.

Jianfa Ma is currently an undergraduate student at the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include the Occupancy Network, Embodied Intelligence, and Computer Vision.

Ziyue Zhao is an undergraduate student majoring in Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include autonomous driving, computer vision.

Qining Qi is an undergraduate student majoring in Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include autonomous driving, computer vision, and information security.

Yechong Liu is a postgraduate student majoring in Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. His research interests include autonomous driving, computer vision, and robust learning.

Qi Hao (Senior Member, IEEE) is currently a full Professor at the Southern University of Science and Technology, Shenzhen, China, the leader of the Intelligent Sensing and Unmanned Systems Lab, Executive Dean of the Research Institute of Trustworthy Autonomous Systems, Executive Director of SUSTech-Haylion Center for Intelligent Transportation, Deputy Director of Shenzhen Key Laboratory of Robotics and Computer Vision. He received the B.E. and M.E. degrees from Shanghai Jiao Tong University, Shanghai, China, in 1994 and 1997, respectively, and the Ph.D. degree from Duke University, Durham, NC, USA, in 2006, all in electrical and computer engineering. From 2007 to 2014, he was an Assistant Professor at the Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL, USA. His current research interests include intelligent unmanned systems, machine learning, and smart sensors.