

Gaussian World Model for Streaming 3D Occupancy Prediction

Supplementary Material

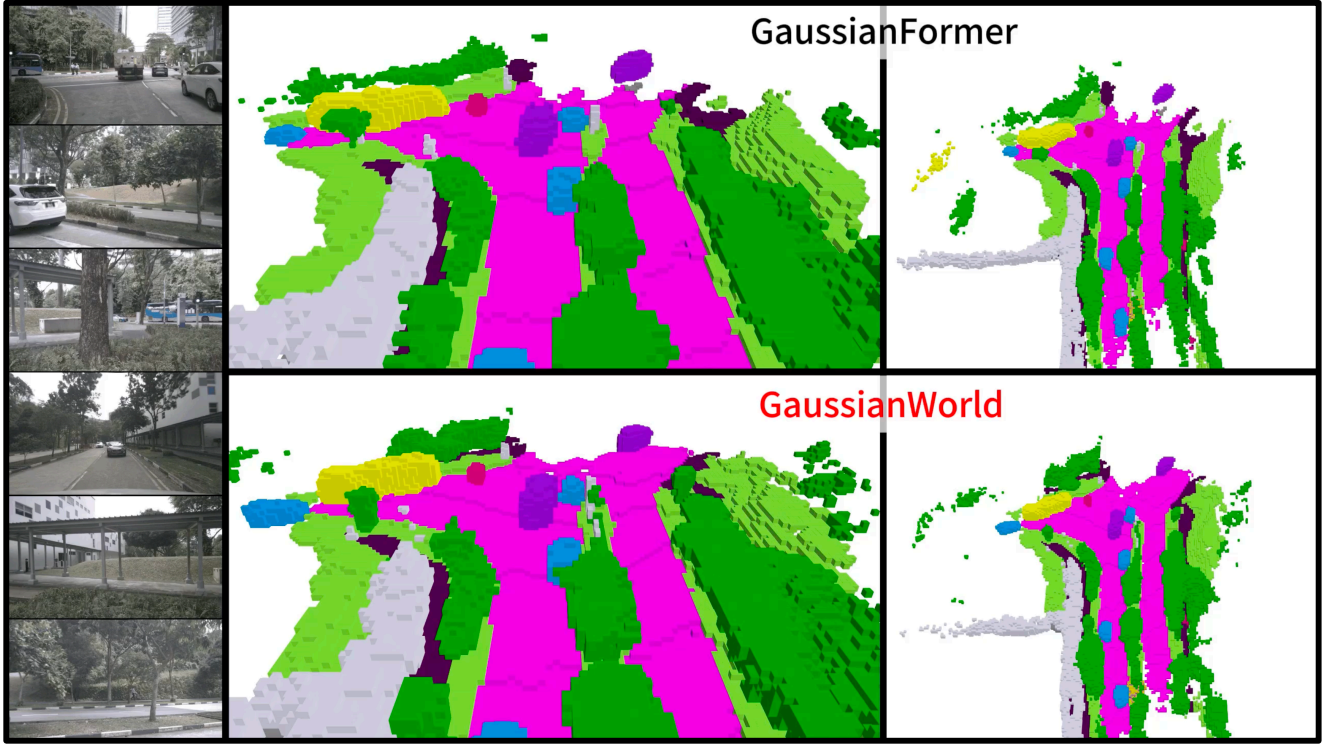


Figure 1. **Visualizations of the proposed GaussianWorld compared to GaussianFormer [2] for 3D semantic occupancy prediction on the nuScenes [1] validation set.** We visualize the six surrounding camera inputs and the corresponding occupancy prediction results. The upper row shows the predicted occupancy by GaussianFormer in the global view(left) and the bird’s eye view(right). The lower row shows the prediction results of GaussianWorld.

A. Additional Implementation Details

Evolution Layer. We employ a unified evolution layer to model the progression of aligned historical Gaussians and the perception of newly completed Gaussians. Following [2], we first voxelize all 3D Gaussians and utilize a 3D sparse convolution block to facilitate interaction between Gaussians. We adopt deformable attention to enable interaction between the Gaussians and image features. Finally, we use a unified refinement block to update the properties of historical Gaussians and new Gaussians separately.

Unified Refinement Block. In this module, a unified prediction layer is employed to predict the property modifications for all Gaussians. For newly completed Gaussians, the predicted changes are directly incorporated into their original properties. For historical Gaussians, only the positions of the dynamic Gaussians are updated. This is accomplished by using the probability of the Gaussian’s dynamic semantic category as semantic weights, which are multiplied by the positional changes before updating the position of Gaussians.

Refinement Layer. To address the misalignment between the 3D Gaussian representation and the real world, we also employ a refinement layer to fine-tune all properties of the Gaussians. The only difference from the evolution layer is that we use an additional temporal weight to update all properties of historical Gaussians

B. Additional Experiments

We have refined the model architecture of GaussianFormer [2] with several key modifications. To learn the scene evolution, we introduce an additional temporal feature attribute to capture the historical information of 3D Gaussians. Rather than predicting the updated properties directly, we predict the changes to Gaussian properties, thereby preserving their original characteristics as much as possible while modeling the scene evolution. Considering that Gaussians need a broad range of movement to model dynamic object motion, we expand the influence range of Gaussians when interacting with images and predicting occupancy. We conduct ablation studies to validate the effec-

040 tiveness of these designs. As shown in Table 1, the absence
041 of these designs results in a slight performance degradation.

Table 1. **Ablation on the model structure design.** Temp. Feat., Delta Ref., and Range denote utilizing the additional temporal feature property, predicting the changes to Gaussian properties, and expanding the influence range of Gaussians, respectively.

Temp. Feat.	Delta Ref.	Range	mIoU	IoU
	✓	✓	21.55	32.81
✓		✓	21.37	32.25
✓	✓		21.21	32.32
✓	✓	✓	21.87	33.02

042 **C. Video Demonstration**

043 Figure 1 shows a sampled image from the video demo for
044 3D semantic occupancy prediction on the nuScenes [1] val-
045 idation set. Compared to GaussianFormer [2], our Gaus-
046 sianWorld shows more cross-frame consistency, especially
047 for static elements. This demonstrates the effectiveness of
048 our Gaussian-based explicit streaming modeling.

049 **References**

050 [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora,
051 Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Gi-
052 ancario Baldan, and Oscar Beijbom. nuscenes: A multimodal
053 dataset for autonomous driving. In *CVPR*, 2020. 1, 2
054 [2] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou,
055 and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-
056 based 3d semantic occupancy prediction. *arXiv preprint*
057 *arXiv:2405.17429*, 2024. 1, 2