GaussianFormer: Scene as Gaussians for Vision-Based 3D Semantic Occupancy Prediction Supplementary Material

Yuanhui Huang[1], Wenzhao Zheng[1,2*], Yunpeng Zhang[3], Jie Zhou[1], and Jiwen Lu[1†]

[1]Tsinghua University [2]University of California, Berkeley [3]PhiGent Robotics
https://wzzheng.net/GaussianFormer
huangyh22@mails.tsinghua.edu.cn; wenzhao.zheng@outlook.com;
yunpengzhang97@gmail.com; {jzhou,lujiwen}@tsinghua.edu.cn

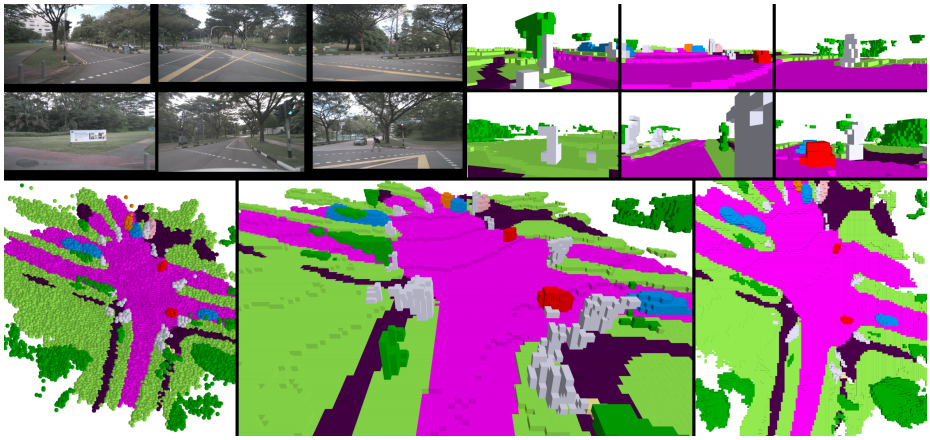

Fig. 1: Visualizations of the proposed GaussianFormer method for 3D semantic occupancy prediction on the nuScenes [1] validation set. We visualize the six input surrounding images and the corresponding predicted semantic occupancy in the upper part. The lower row shows the predicted 3D Gaussians (left), the predicted semantic occupancy in the global view (middle) and the bird's eye view (right).

A Video Demonstration

Fig. 1 shows a sampled image from the video demo[1] for 3D semantic occupancy prediction on the nuScenes [1] validation set. GaussianFormer successfully predicts both realistic and holistic semantic occupancy results with only surround images as input. The similarity between the 3D Gaussians and the predicted occupancy suggests the expressiveness of the efficient 3D Gaussian representation.

* Project Leader
† Corresponding Author
[1] https://wzzheng.net/GaussianFormer

Table 1: Analysis on different design choices.

Gaussian Number	Photometric	Split & Prune	Initialization	mIoU	IoU
144000	×	×	learnable	19.10	29.83
192000	×	×	learnable	19.65	30.37
256000	×	×	learnable	**19.76**	**30.51**
51200	×	×	learnable	16.41	29.37
51200	✓	×	learnable	16.24	**29.43**
51200	×	✓	learnable	**16.50**	29.41
51200	×	×	uniform	16.27	29.23
51200	×	×	pseudo points	18.99	28.84
51200	×	×	G.T. points	**26.78**	**41.81**

B Additional Ablation Study

We conduct additional ablation study on the number of Gaussians, photometric supervision, splitting & pruning strategy, and initialization schemes in Table 1. We observe consistent improvement as the number of Gaussians increases.

We also experiment with an additional photometric loss to reconstruct the input image in Table 1, which does not demonstrates a significant improvement. This is because using photometric loss on nuScenes where the surrounding cameras share little overlap view cannot provide further structural information. In addition, we experiment with the splitting and pruning strategy prevalent in the offline 3D-GS literature and observe that it improves performance compared with the baseline.

We employ learnable properties as initialization for our online model to learn a prior from different scenes in the main paper. We further experimented with different initialization strategies including uniform distribution, predicted pseudo point cloud, and ground-truth point cloud. We see that initialization is important to the performance and depth information is especially crucial.

C Additional Visualizations

In Fig. 2, we provide the visualizations of GaussianFormer for 3D semantic occupancy prediction on the KITTI-360 [2] validation set. Similar to the visualizations for nuScenes [1], we observe that the density of the 3D Gaussians is higher with the presence of vehicles (e.g. the 4th row) which demonstrates the object-centric nature of the 3D Gaussian representation and further benefits resource allocation. In addition to overall structure, GaussianFormer also captures intricate details such as poles in the scenes (e.g. the 1st row). The discrepancy of the density and scale of the 3D Gaussians on nuScenes and KITTI-360 is because we set the number of Gaussians to 144000 / 38400 and the max scale of Gaussians to 0.3m / 0.5m for nuScenes and KITTI-360, respectively. We also visualize the output 3D semantic Gaussians of each refinement layer in Fig. 3.

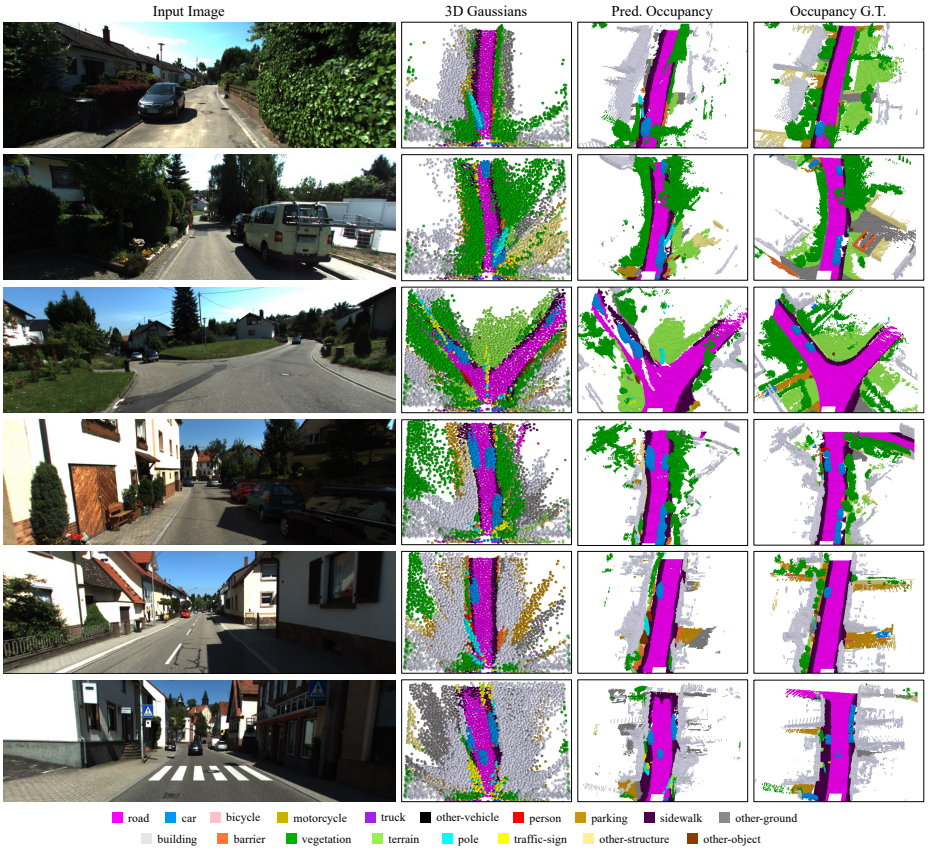

Fig. 2: Visualizations of the proposed GaussianFormer method for 3D semantic occupancy prediction on the KITTI-360 [2] validation set. GaussianFormer is able to capture both the overall structures and intricate details of the driving scenes in a monocular setting.

In Fig. 4, we provide a qualitative comparison between our GaussianFormer and SurroundOcc [3].

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
2. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. PAMI (2022)
3. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: ICCV. pp. 21729–21740 (2023)

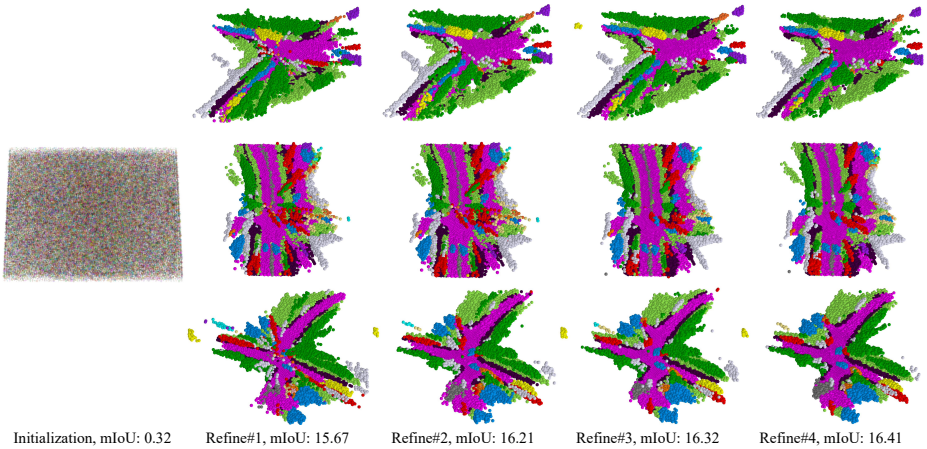

Fig. 3: Visualization of the outputs of the refinement layers and the corresponding mIoU on nuScenes.

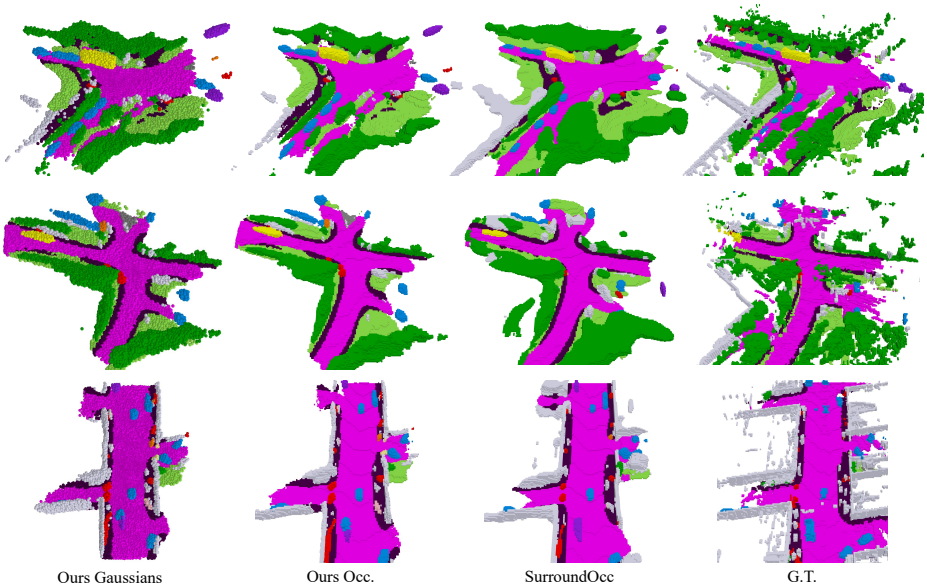

Fig. 4: Qualitative comparison between our method and SurroundOcc on nuScenes.