GaussianFormer-2: Probabilistic Gaussian Superposition for Efficient 3D Occupancy Prediction

Supplementary Material

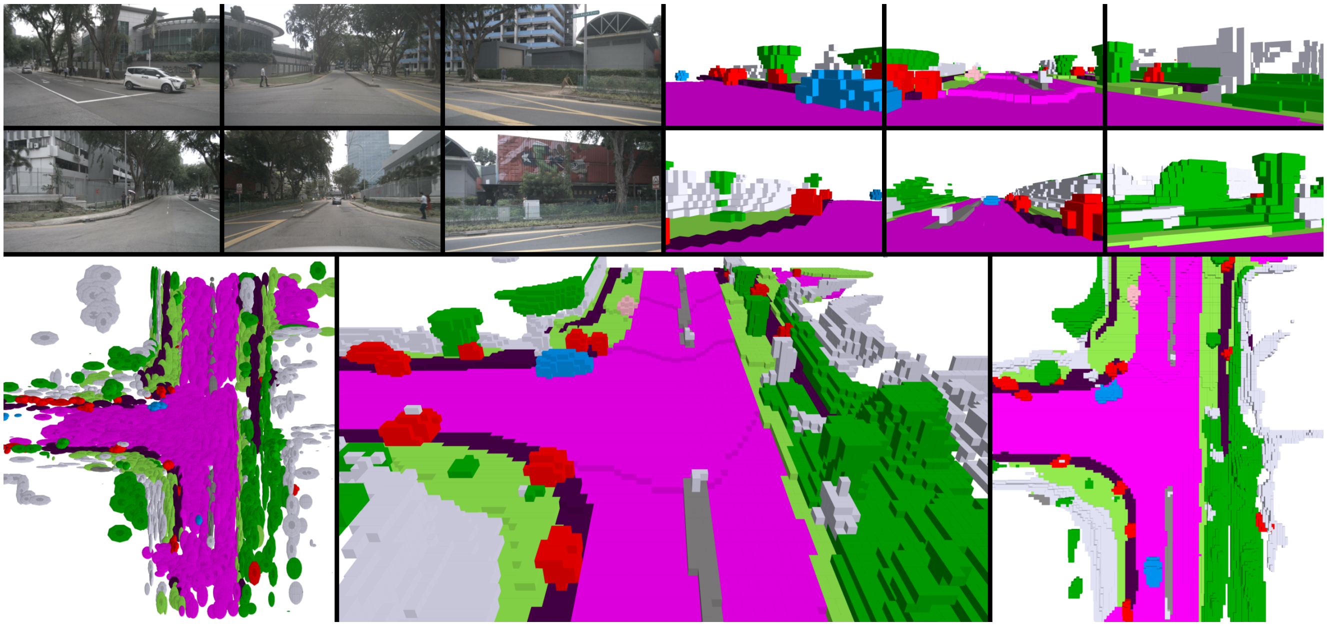

Fig. 1. **Visualizations of Gaussians, camera-view and overall occupancy on nuScenes.** We provide the input RGB images and their corresponding camera-view occupancy in the upper part. And we visualize the predicted 3D Gausians (left), the semantic occupancy in the global view (middle), and in the bird's eye view (right) in the lower part.

A. Video Demonstration

Fig. 1 shows a sampled frame of our video demonstration included in the supplementary material for 3D semantic occupancy prediction on the nuScenes dataset [2]. We note that the camera-view occupancy visualizations align well with the input RGB images. Moreover, each instance is sparsely described by only a few Gaussians with adaptive shapes, which demonstrates the efficiency and the object-centric nature of our model.

B. Visualizations on KITTI-360

We provide visualization results of Gaussians and occupancy on the KITTI-360 dataset [4] in Fig. 2. We observe that our GaussianFormer-2 is able to predict both intricate geometry and semantics of the 3D scene. Furthermore, the 3D Gaussians in our model are adaptive in their scales according to the specific objects they are describing, compared with isotropic spherical Gaussians with maximum scales in GaussianFormer [3].

C. Comparison with Other Efficient Methods

In this section, we provide a quantitative comparison between the 3D Gaussian representation and other sparse methods in Tab. 1. In summary, sparse-voxel-based methods are limited by predefined grid patterns and cannot represent fine-grained structures. Point-based models assume a homogeneous influence on the neighborhood, resulting in less expressiveness.

Tab 1. **Comparisons with other methods on SurroundOcc.**

Method	Representation	Resolution	mIoU	IoU
SparseOcc [5]	Sparse voxels	32000	16.14	28.20
OPUS [6]	Points	76800	16.67	24.02
GaussianFormer-2	3D Gaussians	**12800**	**20.82**	**31.74**

D. Further Analysis

D.1. Multiplication Theorem of Probability

We explain the effectiveness of the probability multiplication theorem in modeling the geometry structure in Fig. 3. First, the inequality $1 > \alpha(\mathbf{x}) > \alpha(\mathbf{x}; \mathbf{G}_i) > 0$ holds for any Gaussian $\mathbf{G}_i$, which implies the confidence of $\mathbf{x}$ being occupied would be large enough ($\alpha(\mathbf{x}) \rightarrow 1$) if any single Gaussian is close to it ($\alpha(\mathbf{x}; \mathbf{G}_i) \rightarrow 1$). Second, the gradient of $\alpha(\mathbf{x})$ w.r.t. $\alpha(\mathbf{x}; \mathbf{G}_i)$ writes $\prod_{j \neq i}(1 - \alpha(\mathbf{x}; \mathbf{G}_j))$, which produces adaptive gradients according to the different contributions of Gaussians, making each Gaussian focus on its local neighborhood. In contrast, the additive form is

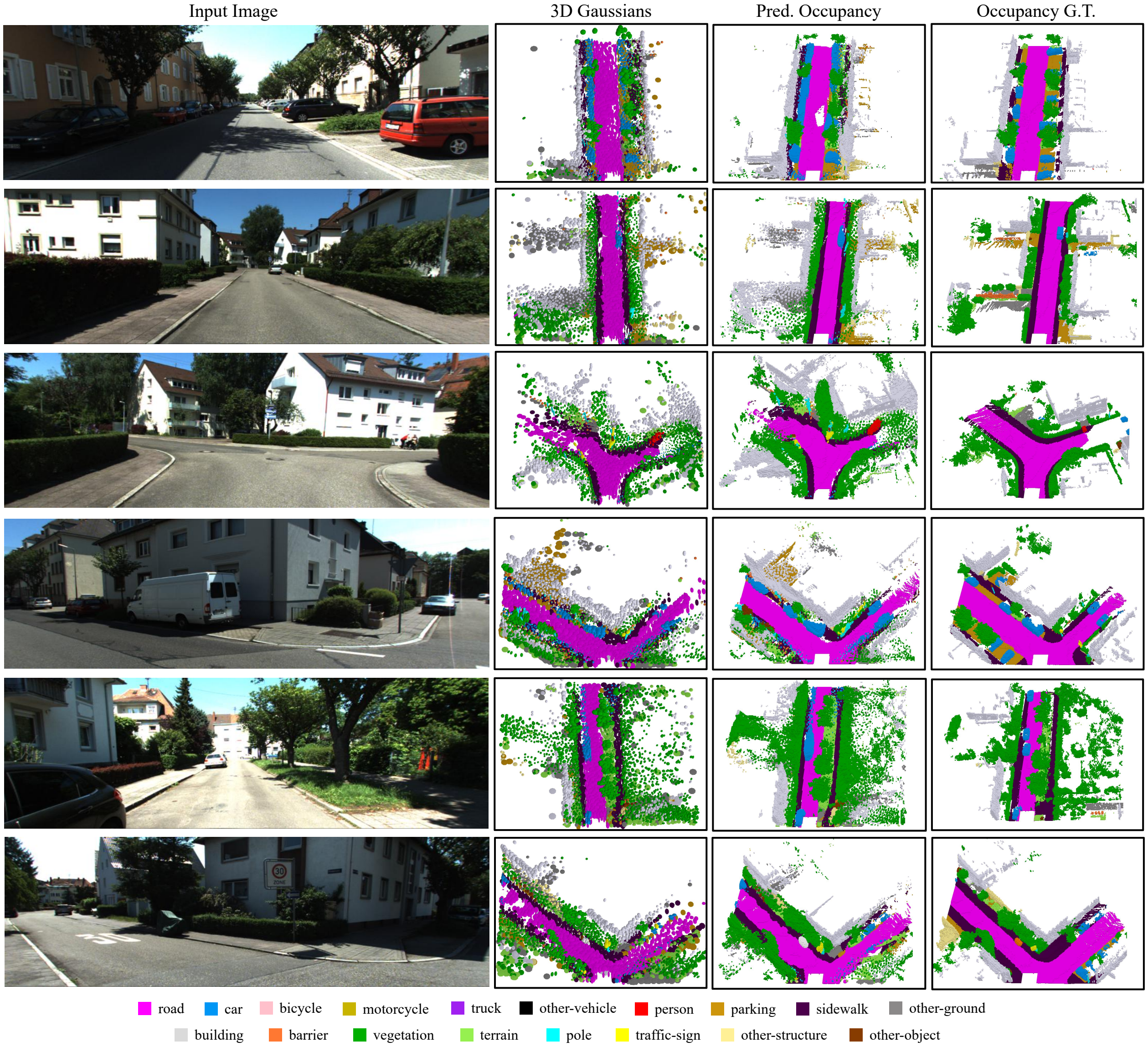

Fig. 2. **Visualizations of Gaussians and occupancy on KITTI-360.** Our method captures both the intricate geometry and semantics of the scene with shape-adaptive Gaussians.

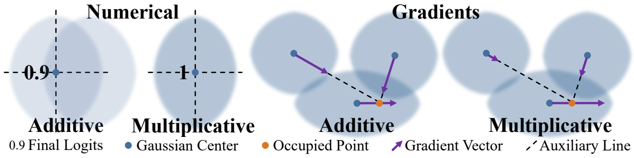

Fig. 3. **Illustration for advantages of the multiplicative form.**

unbounded and equally encourages all Gaussians to overlap with uniform gradients.

D.2. Gaussian Mixture Model for Semantics

Since GaussianFormer does not normalize the semantics, we compare our method with an adapted version of GaussianFormer with semantic normalization to further verify the effectiveness of the probabilistic design. We have observed suboptimal performance in Tab. 2 if we simply normalize semantics without geometric probabilistic modeling.

To provide further analysis on semantic modeling, the final normalized logits for point $\mathbf{x}$ writes $\mathbf{e}(\mathbf{x};\mathcal{G}) = \sum_i p(\mathbf{G}_i|\mathbf{x})\tilde{\mathbf{c}}_i$ as in Eq. (6). Since $\tilde{\mathbf{c}}_i$s are irrelevant to spatial overlap, we can regard them as constants. The cross-entropy loss will maximize $p(\mathbf{G}^*|\mathbf{x})$ corresponding to the largest $\tilde{\mathbf{c}}_*^{y_\mathbf{x}} = \max(\tilde{\mathbf{c}}_1^{y_\mathbf{x}}, ..., \tilde{\mathbf{c}}_P^{y_\mathbf{x}})$, where $y_\mathbf{x}$ denote the label of $\mathbf{x}$. Maximizing $p(\mathbf{G}^*|\mathbf{x})$ will then pull closer $\mathbf{G}^*$ and push away other Gaussians.

In addition, the geometry and semantic predictions are one-stage. And we think they are complementary and orthogonal because in essence, geometry and semantic predictions focus on optimizing $\alpha(\mathbf{x};\mathbf{G}_i)$ and $\tilde{\mathbf{c}}_i$, respectively.

Tab 2. **Ablation on semantic normalization.**

Method	Normalize Semantics	mIoU	IoU
GaussianFormer	×	16.00	28.72
GaussianFormer	✓	18.90	29.45
Ours	✓	**20.32**	**31.04**

E. Metric Details

Position. Gaussians, even after full training, can still be found in unoccupied space due to the localized nature of the receptive field. These Gaussians fail to describe meaningful structures, rendering them ineffective and devoid of practical utility. A higher proportion of Gaussians in unoccupied space indicates suboptimal utilization. Hence, we define the *percentage of Gaussians in correct positions (Perc.)* as:

$$\text{Perc.} = \frac{N_{\text{correct}}}{N_{\text{total}}} \cdot 100\%, \tag{1}$$

where N_{correct}, and N_{total} denote the number of Gaussians of which means are in occupied space, and the total number of Gaussians, respectively. A higher percentage indicates a better alignment of the Gaussians with occupied or meaningful area in the space, thus reflecting a more efficient use of the model's capacity.

The above measurement provides a hard evaluation, where Gaussians are either classified as being in correct or incorrect positions without considering their proximity to the nearest occupied area. This binary approach does not capture how close Gaussians in unoccupied regions are to meaningful positions. To address this limitation, we define a complementary soft measurement as the average distance of each Gaussian to its nearest occupied voxel center, denoted as *Dist.* (in meters), computed as follows:

$$\text{Dist.} = \frac{1}{P}\sum_{i=1}^{P} \min_{\mathbf{v}\in\mathcal{V}} ||\mathbf{m}_i - \mathbf{v}||_1, \tag{2}$$

where $\mathbf{m}_i$, $\mathcal{V}$, $\mathbf{v}$, and $||\mathbf{m}_i - \mathbf{v}||_1$ denote the mean of the i-th Gaussian, the set of occupied voxel centers, the center of one voxel in this set, and L1 distance between the mean of the Gaussian and the voxel center, respectively. Note that this distance is calculated with respect to the voxel center, and thus Gaussians positioned within the correct occupied area may also have a non-zero distance.

Overlap. The *overall overlapping ratio of Gaussians (Overall.)* provides a global perspective on the redundancy in the Gaussian representation. Each Gaussian is modeled as an ellipsoid, where the semi-axis lengths are derived from the Mahalanobis distance at a chi-squared value of 6.251, corresponding to the 90% confidence level of a Gaussian distribution in three degrees of freedom (DoFs). The *Overall.* is then calculated as the ratio of the summed 90% confidence volumes $V_{i,90\%}$ of all Gaussians to the total coverage volume of all Gaussians V_{coverage} in the scene:

$$\text{Overall.} = \frac{\sum_{i=1}^{P} V_{i,90\%}}{V_{\text{coverage}}}, \tag{3}$$

where V_{coverage} represents the volume of all Gaussians combined as a unified shape. To estimate V_{coverage}, we employ the *Monte Carlo method* where a large number of points are randomly sampled within the bounding box of the scene. For each sampled point, we check whether it lies within the 90% confidence ellipsoid of any Gaussian. The volume is then approximated as:

$$V_{\text{coverage}} = V_{\text{scene}} \cdot \frac{N_{\text{in}}}{N_{\text{total}}}, \tag{4}$$

where N_{in}, and N_{total} are the number of sampled points that fall within the 90% confidence ellipsoid of at least one Gaussian, and the total number of sampled points, respectively. This approach ensures an accurate estimation of the unified volume, efficiently handling the overlapping regions of the Gaussians by not double-counting them.

We next detail the derivation of the ellipsoid volume corresponding to the 90% confidence region of a 3D Gaussian distribution. Considering a multivariate Gaussian distribution in 3D defined as:

$$\mathbf{g}(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}|^{1/2}} \exp\big(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\big), \tag{5}$$

where $\mathbf{x}$, $\mathbf{\Sigma}$, and $|\mathbf{\Sigma}|$ are the mean vector, 3x3 covariance matrix, and the determinant of the covariance matrix, respectively. The *Mahalanobis distance* d of point $\mathbf{x}$ from the mean $\mathbf{m}$ is defined as:

$$d^2(\mathbf{x},\mathbf{m}) = (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}). \tag{6}$$

The 90% confidence region of the Gaussian distribution corresponds to the set of points for which the Mahalanobis distance satisfies:

$$d^2 \leq \chi^2_{3,0.9} \approx 6.251, \tag{7}$$

where $\chi^2_{3,0.9}$ is the chi-square critical value for three degrees of freedom at the 90% confidence level. For a Gaussian distribution, the semi-axis lengths are determined by the square root of the eigenvalues of $\mathbf{\Sigma}$, scaled by $\chi^2_{3,0.9}$. Thus, the volume of the ellipsoid from 90% of the 3D Gaussian distribution is:

$$V_{90\%} = \frac{4}{3}\pi(6.251)^{3/2}|\mathbf{\Sigma}|^{1/2}. \tag{8}$$

A higher value of *Overall.* indicates greater overlapping volumes among the Gaussians, signifying redundancy in Gaussian representation. This metric provides insights into the utilization of Gaussians to represent the scene.

The *individual overlapping ratio of Gaussians (Indiv.)* offers a fine-grained analysis of the overlap between Gaussians in a scene. This measurement quantifies the degree to which each Gaussian overlaps with all other Gaussians, averaged across all Gaussians in the scene. The value of this metric indicates approximately how many times the volume of a single Gaussian is fully overlapped with other Gaussians on average. To compute this, we use the Bhattacharyya coefficient [1], which measures the similarity between two Gaussian distributions. The *individual overlapping ratio* is defined as:

$$\text{Indiv.} = \frac{1}{P}\sum_{i=1}^{P}\left(\sum_{j\neq i}\text{BC}_{i,j}\right), \tag{9}$$

where $\text{BC}_{i,j}$ is the Bhattacharyya coefficient between the i-th and j-th Gaussians, given by:

$$\text{BC}_{i,j} = \frac{\sqrt[4]{|\mathbf{\Sigma}_i||\mathbf{\Sigma}_j|}}{\sqrt{|\mathbf{\Sigma}_{ij}|}}e^{-\frac{1}{8}(\mathbf{m}_i-\mathbf{m}_j)^{\text{T}}\mathbf{\Sigma}_{ij}^{-1}(\mathbf{m}_i-\mathbf{m}_j)}, \tag{10}$$

where $\mathbf{\Sigma}_{ij} = \frac{1}{2}(\mathbf{\Sigma}_i + \mathbf{\Sigma}_j)$ is the average covariance matrix. A higher value of *Indiv.* indicates more redundancy, as Gaussians are heavily overlapping with each other.

F. Limitations and Failure Cases

We observe that the temporal flickering of Gaussians in the video demonstration is one of the main limitations, and we think streaming prediction considering past frames would alleviate this problem. Furthermore, although the Gaussians in GaussianFormer-2 show a tendency to move towards occupied regions as shown in Figure 7 thanks to the distribution-based initialization, it is still worth investigating how to guide Gaussians more effectively.

References

[1] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society*, 35: 99–110, 1943. 4

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[3] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2405.17429*, 2024. 1

[4] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *PAMI*, 2022. 1

[5] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 1

[6] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Mingming Cheng. Opus: occupancy prediction using a sparse set. In *NIPS*, 2024. 1