Vision-Only Gaussian Splatting for Collaborative Semantic Occupancy Prediction

Cheng Chen[1], Hao Huang[2*], Saurabh Bagchi[1*]

[1]Purdue University
[2]New York University Abu Dhabi
chen4384@purdue.edu, hh1811@nyu.edu, sbagchi@purdue.edu

Abstract

Collaborative perception enables connected vehicles to share information, overcoming occlusions and extending the limited sensing range inherent in single-agent (non-collaborative) systems. Existing vision-only methods for 3D semantic occupancy prediction commonly rely on dense 3D voxels, which incur high communication costs, or 2D planar features, which require accurate depth estimation or additional supervision, limiting their applicability to collaborative scenarios. To address these challenges, we propose the first approach leveraging sparse 3D semantic Gaussian splatting for collaborative 3D semantic occupancy prediction. By sharing and fusing intermediate Gaussian primitives, our method provides three benefits: a neighborhood-based cross-agent fusion that removes duplicates and suppresses noisy or inconsistent Gaussians; a joint encoding of geometry and semantics in each primitive, which reduces reliance on depth supervision and allows simple rigid alignment; and sparse, object-centric messages that preserve structural information while reducing communication volume. Extensive experiments demonstrate that our approach outperforms single-agent perception and baseline collaborative methods by +8.42 and +3.28 points in mIoU, and +5.11 and +22.41 points in IoU, respectively. When further reducing the number of transmitted Gaussians, our method still achieves a +1.9 improvement in mIoU, using only 34.6% communication volume, highlighting robust performance under limited communication budgets.

Code — https://github.com/ChengChen2020/VOGS-CP

Introduction

Multi-agent collaborative perception (CP), also known as cooperative perception, significantly enhances transportation safety, mobility, and efficiency by enabling connected vehicles to integrate multiple viewpoints through Vehicle-to-Everything (V2X) communication, forming a comprehensive understanding of their environment (Huang et al. 2023a). Early CP studies primarily focused on established single-agent tasks such as 3D object detection (3DOD) (Liu et al. 2020; Hu et al. 2022; Ding et al. 2025; Zhang et al. 2024; Xu et al. 2022b, 2023) and 2D bird's-eye-view (BEV)
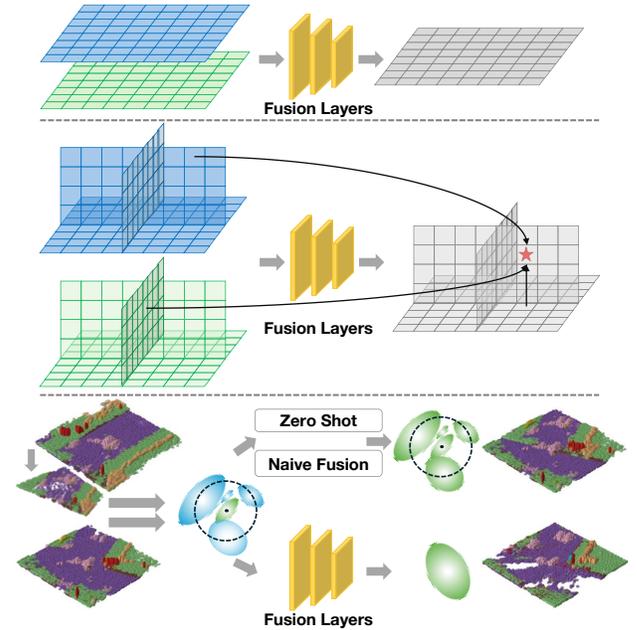
Figure 1: Comparison of shared representations. BEV and tri-plane methods transmit implicit planar features as messages, losing geometric detail and complicating alignment. We instead share explicit, interpretable 3D Gaussian primitives that preserve 3D structure and enable straightforward cross-agent fusion.

semantic segmentation (Wang et al. 2020; Li et al. 2021), typically by fusing latent BEV features (Figure 1 top). Historically reliant on LiDAR point clouds for their precise geometric measurements, recent advances in vision-only methods have shown competitive performance. For example, CoCa3D (Hu et al. 2023) demonstrates that camera-based 3DOD can match or surpass LiDAR-based approaches, highlighting cameras as viable, cost-effective sensors for complex 3D scene understanding.

However, tasks like 3DOD and BEV segmentation simplify the perception of the scene by omitting crucial 3D semantic details, and BEV-based features inherently suffer from information loss due to height compression. This

limitation hinders holistic scene understanding and downstream decision-making. 3D semantic occupancy prediction (SOP) (Roldao, De Charette, and Verroust-Blondet 2022; Tian et al. 2023; Li et al. 2024) bridges this gap by predicting the occupancy status of each voxel in the surrounding 3D space, thus delivering fine-grained semantic and geometric scene information. Vision-centric SOP methods (Cao and De Charette 2022; Song et al. 2017; Chen et al. 2020; Li et al. 2023; Wei et al. 2023; Huang et al. 2023b) have shown promising results on large-scale datasets, yet extending SOP to collaborative scenarios remains largely unexplored. CoHFF (Song et al. 2024) introduced the first collaborative SOP framework, demonstrating the benefits of V2X feature fusion. Nevertheless, it relies on planar tri-plane features (Huang et al. 2023b; Fridovich-Keil et al. 2023) (Figure 1 middle) which requires explicit depth supervision, and employs complex multi-stage, multi-task training pipelines, which introduces computational inefficiencies and complicates cross-agent feature alignment. Inspired by advances in 3D Gaussian splatting (Kerbl et al. 2023), sparse, object-centric representations have emerged as promising alternatives to traditional dense voxel or planar encodings for SOP (Li et al. 2023; Wei et al. 2023; Huang et al. 2023b). Driving scenes are highly sparse—most voxels are empty—so anisotropic Gaussians can shape and orient along semantic object surfaces, concentrating capacity where geometry and semantics lie while keeping free space compact, yielding an efficient 3D scene model. GaussianFormer (Huang et al. 2024b,a) represents scenes using sets of 3D Gaussians, each defined by a mean, covariance, and semantic label; these Gaussians are refined via deformable attention and rendered to a voxel grid through Gaussian-to-voxel splatting.

In this work, we propose the first Vision-Only Gaussian Splatting framework for collaborative SOP that can be efficiently trained end-to-end in a single stage. Our method leverages sparse 3D semantic Gaussians as intermediate representations shared among vehicles, explicitly encoding both geometric and semantic information. By transmitting rigid transforms of Gaussians only within each ego vehicle's region of interest (Figure 1 bottom), our approach effectively integrates multi-agent viewpoints, overcoming occlusions. Additionally, we introduce a novel cross-agent Gaussian fusion module that refines noisy, redundant, and conflicting Gaussians through neighborhood-based fusion, enabling accurate and realistic semantic occupancy prediction. To the best of our knowledge, we are the first to utilize 3D Gaussian splatting for multi-agent collaborative 3D semantic occupancy prediction. We summarize our contributions as:

- We propose the first vision-only Gaussian splatting framework for collaborative SOP, which aggregates 3D Gaussians across agents, improving robustness to occlusions and viewpoint fragmentation.
- We employ a learned neighborhood-based fusion module specifically designed to reduce noise and inconsistencies across multi-agent predictions.
- Extensive experiments validate our framework's effectiveness and communication efficiency compared to single-

agent and existing collaborative SOP methods, achieving robust performance in downstream tasks such as BEV semantic segmentation.

Related Work

3D Semantic Occupancy Prediction. Semantic occupancy prediction (SOP), also known as semantic scene completion (SSC), estimates per-voxel occupancy together with semantic categories, yielding a volumetric representation of geometry and semantics. Since Tesla described an occupancy network for Full Self-Driving (Tesla 2022), research on 3D occupancy for autonomous driving has steadily accelerated. Methods differ by sensing modality: some use LiDAR point clouds (Zhao et al. 2025) or 4D radar tensors (4DRTs) (Ding et al. 2024), while many recent systems are vision-only and lift image features into 3D using learned geometric priors (Cao and De Charette 2022; Song et al. 2017; Chen et al. 2020; Li et al. 2020; Huang et al. 2023b, 2024b; Wei et al. 2023; Zhang, Zhu, and Du 2023; Li et al. 2023). Most prior work addresses single-vehicle perception. Extending SOP to multi-agent collaborative perception centers on two design decisions: the communication representation and the fusion strategy for processing aggregated information under bandwidth limits and field-of-view differences. Collaborative SOP is only beginning to be explored; CoHFF (Song et al. 2024) reports clear benefits from sharing but relies on multi-stage training and depth estimation, which increases system complexity and complicates cross-agent feature alignment.

Collaborative Perception. Collaborative perception for connected autonomous vehicles (CAVs) uses V2X communication and data fusion to improve scene understanding. Fusion is commonly grouped into *early*, *intermediate*, and *late* (Huang et al. 2023a). Early fusion shares raw sensor data (e.g., point clouds), which is bandwidth demanding and raises privacy concerns (He et al. 2021). Late fusion exchanges object lists, which is lightweight but discards fine details (Shi et al. 2022; Zhu et al. 2024). Most recent systems adopt *intermediate fusion*, which exchanges latent features and retains more task signal while avoiding raw data transfer (Wang et al. 2020; Hu et al. 2022; Xu et al. 2022a,b, 2024; Ding et al. 2025). While many works target LiDAR or hybrid setups for 3D detection or BEV segmentation (Xu et al. 2023; Ding et al. 2025; Zhang et al. 2024; Zimmer et al. 2024; Li et al. 2021; Hu et al. 2022; Yang et al. 2023), vision-only collaboration is increasingly practical for large-scale deployment. CoHFF brought *collaborative semantic occupancy prediction* into focus by demonstrating SOP benefits from V2X feature sharing and fusion (Song et al. 2024).

3D Gaussians for Collaboration. 3D Gaussian Splatting (3DGS) represents scenes with anisotropic Gaussians and renders them by splatting for novel view synthesis (Kerbl et al. 2023). Subsequent work adapts Gaussian splatting to perception tasks, including object detection (Cao, Jv, and Xu 2024; Yan, Zheng, and Duan 2024). For semantic occupancy, the GaussianFormer family splats Gaussians into voxels (Huang et al. 2024b,a). Prior work treats Gaussians

as an *internal* single-agent representation. We instead introduce Gaussians as the *communication* medium for collaboration. This choice provides compact messages, closed-form rigid alignment across agents, region-of-interest culling, and explicit geometry that reduces reliance on separate depth supervision. Our method builds communication and cross-agent fusion for Gaussian primitives.

Proposed Approach

Problem Formulation

We first define a general collaborative perception pipeline at the feature level. Let $\mathcal{S}$ be the set of N connected autonomous vehicles (CAVs) within a communication range δ. For the ith agent in the set $\mathcal{S}$, we denote $\mathbf{O}_i$ as its observation, $f_{\text{enc}}(\cdot)$ as its perception encoder, $f_{\text{fusion}}(\cdot)$ as a fusion sub-network, $f_{\text{head}}(\cdot)$ as its task-specific head layers, and $\mathbf{B}_i$ as the corresponding task-specific output. Then, the collaborative perception network of the ith agent works as follows:

$$\mathbf{F}_i = f_{\text{enc}}(\mathbf{O}_i), \quad i \in \mathcal{S} \ , \tag{1}$$

$$\mathbf{F}_{j \to i} = \Gamma_{j \to i}(\mathbf{F}_j), \quad j \in \mathcal{S} \ , \tag{2}$$

$$\mathbf{H}_i = f_{\text{fusion}}(\{\mathbf{F}_{j \to i}\}_{j \in \mathcal{S}}) \ , \tag{3}$$

$$\mathbf{B}_i = f_{\text{head}}(\mathbf{H}_i) \ , \tag{4}$$

where $\mathbf{F}_i$ is the initial intermediate representation from the ith agent's encoder, $\Gamma_{j \to i}$ is an operator that performs spatial alignment and transmits jth agent's representation, $\mathbf{F}_{j \to i}$ is the spatially aligned representation of jth agent's observation in ith agent's coordinate frame, $\mathbf{H}_i$ is the fused aligned representations from the all neighbor agents in the set $\mathcal{S}$.

For vision-only semantic occupancy prediction, observation $\mathbf{O}_i$ is the RGB images captured by multiple surrounding cameras mounted on the ith agent, and $\mathbf{B}_i$ is the holistic surrounding environment represented as a 3D voxels with one-hot embedding, *i.e.*, $\mathbf{B}_i \in \mathbb{R}^{X \times Y \times Z \times C}$ where X, Y and Z are voxel grid dimensions and C denotes semantic classes. Let $\hat{\mathbf{B}}_i$ represent the ground-truth of the semantic voxels, and Φ represent the collaborative perception network (including the encoder, fusion sub-network, and head layers) parametrized by θ, the objective of collaborative semantic occupancy prediction is defined as follows:

$$\max_\theta \sum_i g\left(\Phi_\theta(\mathbf{O}_i, \{\mathbf{F}_{j \to i}\}_{j \in \mathcal{S}}), \hat{\mathbf{B}}_i\right), \ \text{s.t.} \ \sum_j |\mathbf{F}_{j \to i}| \le \beta \ , \tag{5}$$

where $g(\cdot, \cdot)$ is the metric for optimization. For the semantic occupancy prediction task, we adopt IoU and mIoU. The size of transmitted messages is constrained by a communication budget upper bound $\beta \in \mathbb{R}^+$.

Considering practical bandwidth limits, extending voxel-based single-agent SOP methods (e.g.,(Wei et al. 2023; Li et al. 2023)) to collaboration is not viable: transmitting dense voxel features in $\mathbb{R}^{X \times Y \times Z \times D}$ (where D is the hidden dimension) is too costly. Drawn inspiration from TPV-Former (Huang et al. 2023b), CoHFF proposes to transmit features from orthogonal planes $\mathbf{P}^{xz}$ and $\mathbf{P}^{yz}$, with features $\mathbf{F}^{\mathbf{P}^{xz}} \in \mathbb{R}^{X \times Z \times F}$ and $\mathbf{F}^{\mathbf{P}^{yz}} \in \mathbb{R}^{Y \times Z \times F}$. This reduces the communication volume from $XYZD$ to $(XZ + YZ)D$ and makes collaborative SOP feasible.

However, plane-based features (for example, BEV or tri-planes) do not encode explicit depth or full 3D geometry, so they need extra supervision. TPVFormer uses sparse semantic LiDAR labels, whereas CoHFF trains a separate depth estimation network. In addition, CoHFF adopts a two-stage schedule: the occupancy predictor and the semantic segmenter are trained first, followed by the fusion model. This increases training cost and reduces deployment scalability.

Scene as 3D Gaussian Representation

Inspired by GaussianFormer (Huang et al. 2024b), we opt to represent a scene with a set of 3D Gaussian primitives $\mathcal{G} = \{\mathbf{G}_i \in \mathbb{R}^d | i = 1, \ldots, P\}$. Distinct from Gaussian-Former, we make the set of 3D Gaussian primitives serve as a communication medium for collaborative perception. Specifically, each $\mathbf{G}_i$ describes a local region with its mean $\mathbf{m}_i \in \mathbb{R}^3$, scale $\mathbf{s}_i \in \mathbb{R}^3$, rotation $\mathbf{r}_i \in \mathbb{R}^4$, opacity $a_i \in \mathbb{R}^1$, and semantics $\mathbf{c}_i \in \mathbb{R}^{|\mathcal{C}|}$. These primitives are interpreted as local semantic Gaussian distributions which contribute to the overall occupancy prediction for any given 3D location $\mathbf{x} \in \mathbb{R}^3$ through additive aggregation:

$$\hat{\mathbf{o}}(\mathbf{x}; \mathcal{G}) = \sum_{i=1}^P \mathbf{g}_i(\mathbf{x}; \mathbf{m}_i, \mathbf{s}_i, \mathbf{r}_i, a_i, \mathbf{c}_i) \ , \tag{6}$$

where $\mathbf{g}_i(\mathbf{x}; \cdot)$ denotes the contribution of the ith semantic Gaussian primitive to $\hat{\mathbf{o}}(\mathbf{x}; \mathcal{G})$:

$$\mathbf{g}(\mathbf{x}; \mathbf{G}) = a \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\mathsf{T} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{m})\right) \mathbf{c} \ , \tag{7}$$

$$\boldsymbol{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{R}^\mathsf{T} \ , \ \mathbf{S} = \text{diag}(\mathbf{s}) \ , \ \mathbf{R} = \text{q2r}(\mathbf{r}) \ , \tag{8}$$

where $\boldsymbol{\Sigma}, \mathbf{R}, \mathbf{S}$ represent the covariance matrix, the rotation matrix constructed from the quaternion $\mathbf{r}$ with function q2r$(\cdot)$, and the diagonal scale matrix from function diag$(\cdot)$.

From a single-agent perspective, the pipeline first estimates 3D Gaussian primitives with an image-to-Gaussian module, then renders semantic occupancy via Gaussian-to-voxel splatting. We adopt both components from Gaussian-Former (Huang et al. 2024b). We use 3D Gaussian primitives as a compact communication interface for collaborative perception. A *Gaussian packaging* module transforms each neighbor's Gaussians into the ego frame and culls them to the ego region of interest, reducing the transmitted data. A *cross-agent Gaussian fusion* module then refines the ego set by aggregating consistent multi-view evidence and down-weighting noisy or low-opacity inputs. Collaborative semantic occupancy is obtained by splatting the refined ego Gaussians. The system is illustrated in Fig. 2.

Gaussian Packaging

Unlike dense voxel grid features (Wei et al. 2023; Li et al. 2023) or planar features (Song et al. 2024; Huang et al. 2023b), Gaussian primitives model ellipsoidal probability densities in $\mathbb{R}^3$, which remain well defined under rigid transforms (Kerbl et al. 2023; Huang et al. 2024b), and thus cross-agent 3D Gaussian alignment and transmission are straightforward and interpretable. Specifically, let the known extrinsic from agent j to agent i be $T_{ij}(x) = U_{ij}x + t_{ij}$
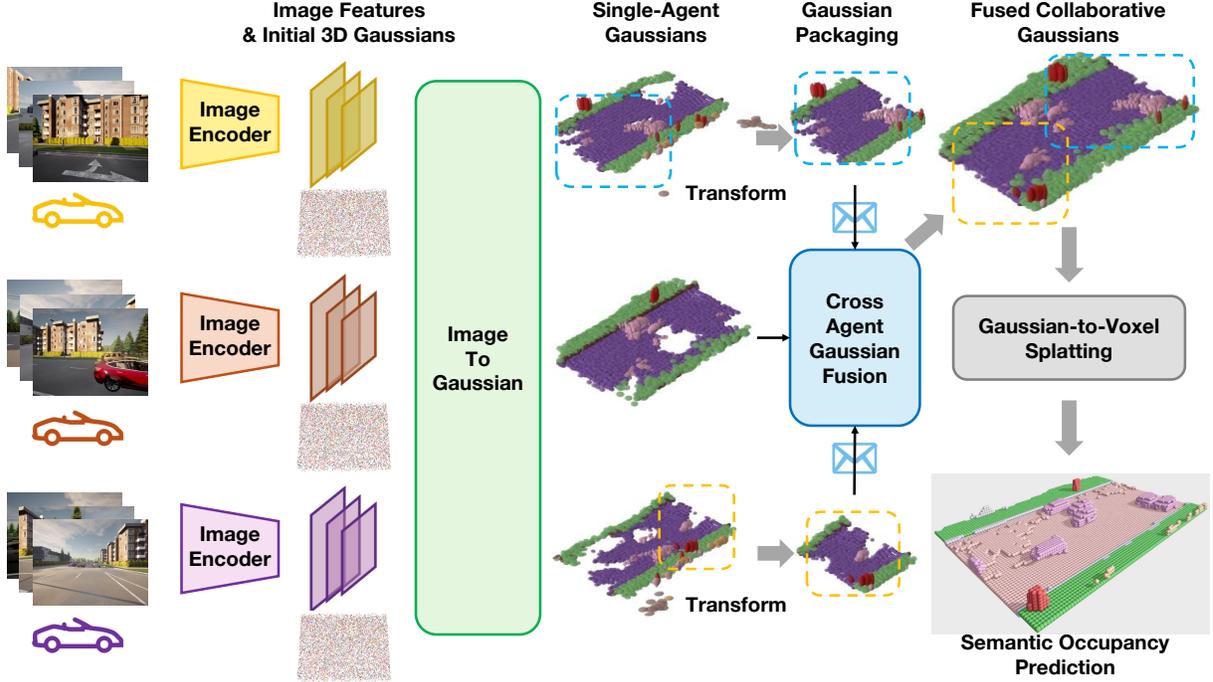
Figure 2: Overview of the proposed pipeline. An initial set of randomly initialized 3D Gaussians is refined by an Image-to-Gaussian module (Huang et al. 2024b) that attends to multi-scale image features, producing single-agent Gaussians. Neighbor agents (top and bottom) are rigidly transformed into the ego frame (middle) and culled to the ego region of interest; the blue and yellow dashed box marks the Gaussians that lie within the ego ROI and are packaged and transmitted to the ego. A cross-agent Gaussian fusion module aggregates these with the ego set. The fused Gaussians are then rendered to semantic occupancy via Gaussian-to-voxel splatting (Huang et al. 2024b). For clarity, the figure shows the *zero-shot* variant.

with $U_{ij} \in \mathrm{SO}(3)$, $t_{ij} \in \mathbb{R}^3$, and quaternion q_{ij} such that q2r$(q_{ij}) = U_{ij}$.

Rigid Transform of a Gaussian. If $X \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ in jth coordinate frame, then $Y = T_{ij}(X)$ is Gaussian in i'th with:

$$\mathbf{m}' = U_{ij}\,\mathbf{m} + t_{ij} \ , \tag{9}$$

$$\boldsymbol{\Sigma}' = U_{ij}\,\boldsymbol{\Sigma}\,U_{ij}^\top = \left(U_{ij}\mathbf{R}\right)\mathbf{S}\mathbf{S}^\top\left(U_{ij}\mathbf{R}\right)^\top \ . \tag{10}$$

Equations (9)–(10) imply that a rigid transform rotates the ellipsoid but does not change its axis lengths. In $(\mathbf{m}, \mathbf{s}, \mathbf{r}, a, \mathbf{c})$ form, we have:

$$\mathbf{m}' = U_{ij}\mathbf{m} + t_{ij} \ , \ \ \mathbf{s}' = \mathbf{s} \ ,$$
$$\mathbf{r}' = q_{ij} \otimes \mathbf{r} \ , \ \ a' = a \ , \ \ \mathbf{c}' = \mathbf{c} \ , \tag{11}$$

where $\otimes$ is quaternion multiplication. Because r and $-r$ encode the same rotation, one may fix a sign convention (*e.g.*, non-negative scalar part) without affecting the transform.

Transmission of a Gaussian. The jth agent only transmits Gaussians whose *transformed* means fall inside the ith agent's region of interest $\mathbf{ROI}_i$ (for example, a 3D volume centered at agent i):

$$\mathcal{G}_{j \to i} = \left\{ \mathbf{G} \in \mathcal{G}_j \mid \mathbf{m}' \in \mathbf{ROI}_i \right\} \ . \tag{12}$$

which greatly reduces the payload compared with transmitting all Gaussians $|\mathcal{G}_j|$ from the jth agent. On reception, the ith agent stacks both its own and the received Gaussians in the same coordinate frame:

$$\mathcal{G}_i^{\mathrm{stack}} = \mathcal{G}_i \cup \{\mathcal{G}_{j \to i}\}_{j \in \mathcal{S}} \ , \tag{13}$$

and proceeds with downstream processing.

Even without joint training, *zero-shot* stacking—using single-agent weights—improves collaborative perception; collective end-to-end training yields further gains (see *Zero Shot* and *Naive Fusion* in Table 1), demonstrating the benefits of an explicit, interpretable representation.

Cross-Agent Gaussian Fusion

Despite the fact that stacking transformed Gaussians improves the performance of collaborative perception for CAVs in semantic occupancy prediction, single agent predictions could be noisy due to occlusions, and different agents may output redundant or inconsistent primitives (Figure 3 last row). We therefore propose a light-weight learnable refinement module that update the ego Gaussians via neighborhood interaction and aggregation.

Unlike the refinement step in the image-to-Gaussian module of GaussianFormer (Huang et al. 2024b), our fusion module decodes neighborhood-conditioned proposals, pools them across nearby Gaussians to suppress noise and enforce consistency, and then blends the pooled update with the ego Gaussian attributes.

Method	Single-Agent		Collaborative Perception			
	CoHFF	GSFormer	CoHFF	Zero Shot	Naive Fusion	Learned Fusion
IoU (↑)	38.52	**67.76**	50.46	67.88	<u>70.10</u>	**72.87**
mIoU (↑)	24.81	**29.20**	34.16	30.54	<u>36.02</u>	**37.44**
Building	**21.04**	3.84	**25.72**	3.91	7.18	<u>9.61</u>
Fence	**20.05**	14.10	<u>27.83</u>	17.29	25.66	**29.20**
Terrain	43.93	68.97	48.30	72.00	**76.05**	<u>74.51</u>
Pole	**31.66**	5.94	**42.74**	8.44	<u>12.67</u>	12.19
Road	55.83	**79.37**	61.77	<u>81.35</u>	78.60	**83.05**
Side walk	17.31	**70.55**	39.62	68.91	<u>75.88</u>	**78.22**
Vegetation	**14.49**	12.54	**20.59**	13.72	16.42	<u>20.43</u>
Vehicles	**58.55**	49.25	**63.28**	49.56	57.14	<u>60.49</u>
Wall	**33.30**	30.79	**58.27**	30.49	35.25	<u>36.45</u>
Guard rail	1.54	**15.01**	1.94	20.78	**41.00**	<u>32.50</u>
Traffic signs	0	0	**16.33**	0	6.35	<u>8.26</u>
Bridge	0	0	<u>3.53</u>	0	0	**4.35**

Table 1: Comparison for semantic occupancy prediction. In the collaborative setting, our *naive fusion* and *learned fusion* variants achieve better IoU and mIoU than CoHFF (Song et al. 2024). The best is in **bold** and the second best is <u>underlined</u>.

Neighborhood and Pairwise Features. Let $\mathcal{G}_i = \{\mathbf{G}_k\}_{k=1}^{N_i}$ be the ego Gaussians generated by the ith agent and $\mathcal{G}_{\text{nbgr}} = \bigcup_j \mathcal{G}_{j \to i}$ be received Gaussians (both in ith's coordinate fraome). For each ego $\mathbf{G}_k = (\mathbf{m}_k, \mathbf{s}_k, \mathbf{r}_k, a_k, \mathbf{c}_k)$ we form a radius-ρ neighborhood:

$$\mathcal{G}_{\text{nbgr},k} = \{\, \mathbf{G}_j \in \mathcal{G}_{\text{nbgr}} \mid \|\mathbf{m}_j - \mathbf{m}_k\|_2 \leq \rho \,\} \,. \quad (14)$$

For each pair of Gaussians $\mathbf{G}_k \in \mathcal{G}_i$ and $\mathbf{G}_{k,j} \in \mathcal{G}_{\text{ngbr},k}$, we build a pairwise feature by concatenating the ego attributes with neighboring cues:

$$\mathbf{f}_k^{\text{ego}} = \left[\, \mathbf{m}_k^\top, \ \mathbf{s}_k^\top, \ \mathbf{r}_k^\top, \ a_k, \ \mathbf{c}_k^\top \,\right] \,,$$
$$\mathbf{f}_{k,j}^{\text{rel}} = \left[\, (\mathbf{m}_j - \mathbf{m}_k)^\top, \ (\mathbf{s}_j - \mathbf{s}_k)^\top, \ |\langle \mathbf{r}_j, \mathbf{r}_k \rangle|, \ a_j, \ \mathbf{c}_j^\top \,\right] \,,$$
$$\mathbf{z}_{k,j} = \left[\, \mathbf{f}_k^{\text{ego}} \mid \mathbf{f}_{k,j}^{\text{rel}} \,\right] \,, \quad (15)$$

where the term $|\langle \mathbf{r}_j, \mathbf{r}_k \rangle|$ is the sign-invariant quaternion cosine; opacity a_j and semantics $\mathbf{c}_j$ are used in absolute form.

Pairwise Feature Refinement. A multi-layer perceptron (MLP) maps $\mathbf{z}_{k,j}$ to a refinement proposal from neighbor j on ego Gaussian k:

$$\mathbf{u}_{k,j} \coloneqq \left(\Delta\mathbf{m}_{k,j}, \ \mathbf{s}_{k,j}^\star, \ \mathbf{r}_{k,j}^\star, \ a_{k,j}^\star, \ \mathbf{c}_{k,j}^\star\right) = \text{MLP}(\mathbf{z}_{k,j}) \,, \quad (16)$$

which are, respectively, a residual update to the Gaussian center, a scale proposal, a quaternion rotation proposal, an opacity proposal, and semantic logits.

Aggregation over Neighbors. We aggregate pairwise proposals with either mean pooling or attention pooling. Mean pooling uses uniform weights $w_{k,j} = 1/|\mathcal{G}_{\text{nbgr},k}|$. Attention pooling uses a learned softmax over neighbors:

$$w_{k,j} = \text{softmax}_j\left(\frac{\langle Q\,\mathbf{x}_k^{\text{ego}}, \ K\,\mathbf{x}_{k,j}^{\text{rel}} \rangle}{\sqrt{d}}\right) \,, \quad (17)$$

where Q, K are learned projection layers. The aggregated proposal is:

$$\bar{\mathbf{u}}_k = \sum_{G_j \in \mathcal{G}_{\text{nbgr},k}} w_{k,j}\,\mathbf{u}_{k,j} \,. \quad (18)$$

Update Rules. Let $\bar{\mathbf{u}}_k = (\overline{\Delta\mathbf{m}}_k, \overline{\mathbf{s}^\star}_k, \overline{\mathbf{r}^\star}_k, \overline{a^\star}_k, \overline{\mathbf{c}^\star}_k)$. We update the ego Gaussian as:

$$\hat{\mathbf{m}}_k = \mathbf{m}_k + \overline{\Delta\mathbf{m}}_k \,, \qquad \hat{\mathbf{s}}_k = \overline{\mathbf{s}^\star}_k \,,$$
$$\hat{\mathbf{r}}_k = \overline{\mathbf{r}^\star}_k \,, \qquad \hat{a}_k = \overline{a^\star}_k \,, \quad (19)$$
$$\hat{\mathbf{c}}_k = \alpha_k\,\mathbf{c}_k + (1 - \alpha_k)\,\overline{\mathbf{c}^\star}_k \,,$$

where the semantic blend uses a confidence weight:

$$\alpha_k = \frac{\text{conf}(\mathbf{c}_k)}{\text{conf}(\mathbf{c}_k) + \text{conf}(\overline{\mathbf{c}^\star}_k)} \,, \quad \text{conf}(v) = \max\left(\frac{v}{\mathbf{1}^\top v + \varepsilon}\right) \,. \quad (20)$$

Experiments

Datasets. OPV2V is a large-scale collaborative perception dataset collected in the CARLA simulator (Dosovitskiy et al. 2017) using the OpenCDA autonomous driving simulation framework with Vehicle-to-Vehicle (V2V) communication, with each sample containing multi-sensor data from 2-7 vehicles. The original OPV2V does not provide semantic occupancy labels, so we use the enhanced Semantic-OPV2V released by CoHFF (Song et al. 2024), which replays the simulations with additional semantic LiDARs. Following the procedure in CoHFF, we aggregate the multi-agent ground truth to obtain the collaborative semantic occupancy supervision.

Implementation Details. Following CoHFF (Song et al. 2024), we utilize a $40 \times 40 \times 3.2$ meter detection area with a grid size of $100 \times 100 \times 8$, resulting in a voxel size of $0.4m^3$. Unless otherwise noted, each agent default utilizes $|\mathcal{G}| = P = 25600$ Gaussians as scene representation, and neighborhood ρ is set to $0.4m$ with attention pooling. We allow CAVs to transmit and share Gaussian primitives for cross-agent fusion. Our experiment incorporates the analysis of 12 semantic labels plus an additional empty label. For optimization, we use AdamW (Loshchilov and Hutter 2017) with weight decay 0.01. The learning rate warms up for the first 500 iterations to 2×10^{-4} and then follows a cosine decay. We train for 30 epochs with batch size

Approach	# Agents	Vehicle	Road	Others
CoBEVT	2	46.13	52.41	-
CoHFF	2	47.40	63.36	40.27
Ours	2	**70.25**	**82.69**	**79.37**
CoBEVT	Up to 7	60.40	63.00	-
CoHFF	Up to 7	64.44	57.28	45.89
Ours	Up to 7	**75.30**	**84.96**	**80.19**

Table 2: Comparison of BEV semantic segmentation with IoU (%) for Vehicle, Road, and Others classes.

Metric	CoHFF	# Gaussians	
		25600	6400
CV (MB) (↓)	0.78	1.07	0.27
IoU (↑)	50.46	72.87	72.42
mIoU (↑)	34.16	37.44	36.02

Table 3: Comparison of communication volume.

# Gaussians	Radius ρ	Attn.	mIoU ↑	IoU ↑
6400	0.4		35.50	71.81
6400	0.8		36.02	72.42
25600	0.4		37.01	73.49
25600	0.4	✓	37.44	72.87
25600	0.8		36.81	72.28
25600	0.8	✓	37.06	73.03

Table 4: Ablation over number of Gaussians, neighborhood radius ρ, and pooling method.

8 on a single NVIDIA A100 (40GB). Unless noted otherwise, all components are trained end to end. Following GaussianFormer (Huang et al. 2024b), we use voxel-wise cross-entropy loss together with Lovász-Softmax (Berman, Triki, and Blaschko 2018) loss, which directly targets IoU improvement. The total loss is:

$$L = L_{\text{CE}} + L_{\text{Lovasz}} . \tag{21}$$

Evaluation Metrics. Following (Cao and De Charette 2022; Huang et al. 2023b; Song et al. 2024), we adopt mean Intersection-over-Union (mIoU) and Intersection-over-Union (IoU) to evaluate the performance of semantic occupancy prediction. For evaluations in subsequent applications, following prior work (Song et al. 2024), we report BEV segmentation using 2D IoU by projecting predicted and ground-truth semantic voxels onto the BEV plane and compute IoU for vehicles, roads and other general objects.

Results and Analysis

Table 1 compares semantic occupancy prediction performance across CoHFF and our fusion variants.

Single-Agent Perception. We observe a large IoU gain (+29.24) steming from explicit free-space modeling: one large fixed Gaussian represents empty space across the scene, while the remaining Gaussians model occupied regions. This reduces confusion between empty and occupied voxels and allows the network focus capacity on object surfaces. The model learns the empty-space primitive and the occupied classes jointly, rather than using a separate occupancy network as in CoHFF. We further observe an increase in mIoU, driven chiefly by better performance on ground-level categories such as sidewalk and terrian.

Collaborative Perception. In collaborative scenarios we evaluate three variants of our approach. The ***zero-shot*** variant enables the ego vehicle directly concatenate Gaussians received from its neighbors before occupancy rendering.

The ***naive fusion*** variant trains the entire multi-agent system end-to-end starting from the zero-shot baseline. The ***learned fusion*** variant applies the proposed cross-agent Gaussian fusion pooling to combine neighboring Gaussians.

The ***zero-shot*** variant improves on the single-agent baseline by a small margin, confirming that Gaussian messages convey useful, explicit 3D evidence. However, stacking Gaussians from multiple agents without coordination introduces noise and inconsistency (see Figure 3), which leads to redundant and inaccurate occupancy predictions. When the system is trained end-to-end on top of this baseline (i.e., ***naive fusion***), the resulting fusion variant achieves larger gains and surpasses CoHFF in mIoU, with performance on more classes (e.g., fence, vegetation, vehicles) matching or exceeding that of CoHFF. The ***learned fusion*** variant refines the ego Gaussians by pooling neighbouring Gaussians shared by other agents, leading to measurable accuracy gains across most categories. Notably, the bridge class—missed entirely by the naive-shot variant—improves from 0% to 4.35% IoU. Though the increase is modest, it shows that the learned fusion helps capture structures that are both sparse and difficult to observe from a single viewpoint.

BEV Segmentation. Table 2 illustrates improved IoU for road and other semantic categories. Although the 3D mIoU for vehicle is slightly lower than that of CoHFF, projecting the predictions to the 2D plane yields a better BEV IoU, indicating that our voxel occupancy provides more accurate height estimates when mapping to ground level.

Communication Volume. Communication volume, measured by the size of transmitted messages, is critical for cooperative perception because deployed systems must operate under limited bandwidth. Our framework exchanges variable-length sets of Gaussian primitives instead of fixed-resolution feature maps, so the communication load rises with the number of Gaussians that cover the overlapping regions of interest among connected vehicles. We therefore report the average message size over the evaluation set. As shown in Table 3, the transmitted volume is roughly proportional to the number of Gaussians. By reducing this count, our method attains better occupancy prediction than CoHFF while using only 34.6% bandwidth.

Ablation Study. Table 4 lists an ablation in which we vary three factors: (i) the number of 3-D Gaussians, (ii) the neighbourhood radius ρ, and (iii) the pooling rule (mean versus attention pooling). Increasing the Gaussian count leads
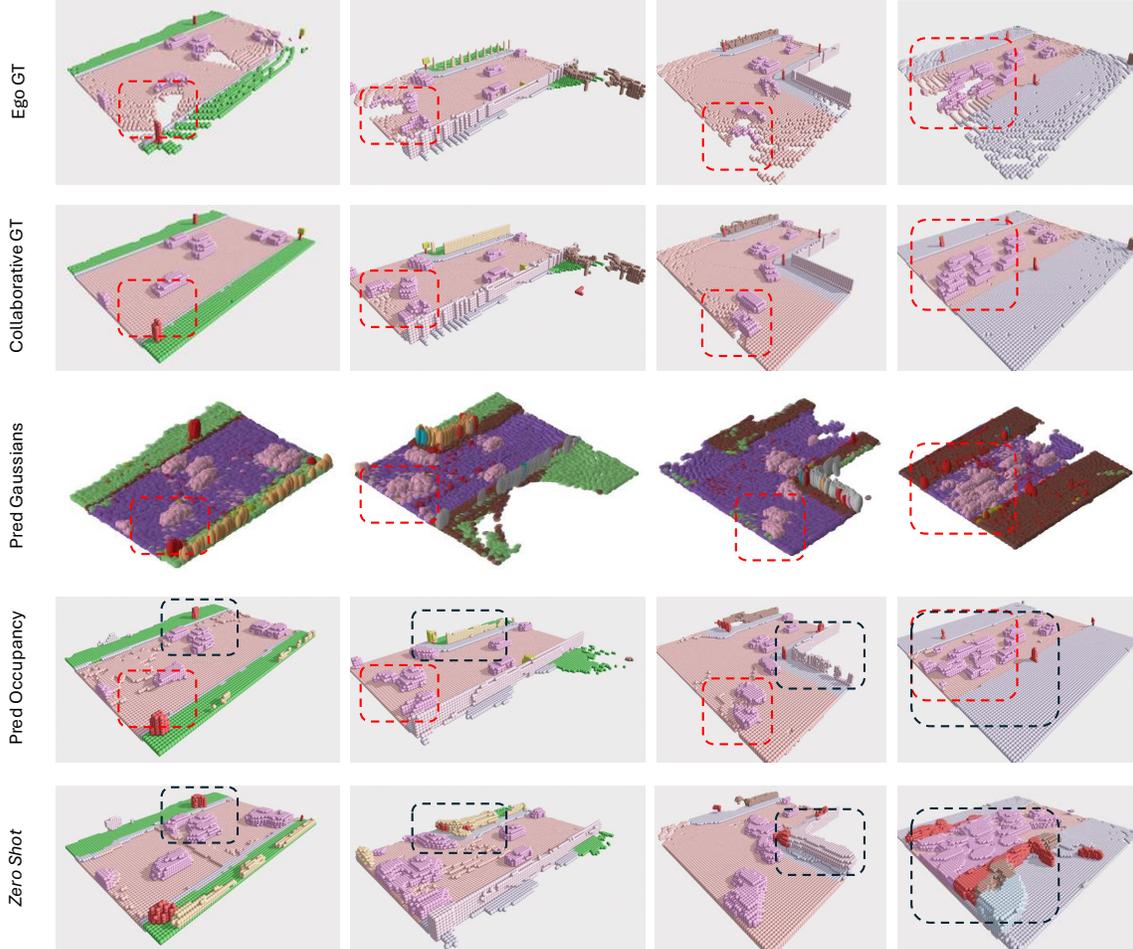
Figure 3: Qualitative comparison of ego-only ground truth, collaborative ground truth, predicted Gaussians, predicted occupancy, and the *zero-shot* variant. Red boxes highlight occluded object structure captured by Gaussian primitives in collaborative setting. The *zero-shot* variant can look plausible but often shows clustered redundancy and noise; black boxes mark cases where the neighborhood-based fusion suppresses redundancy and improves consistency. An opacity threshold is applied for display, so the predicted Gaussians are not exhaustive.

to a rise in mIoU, reflecting the finer geometric detail that a denser set of primitives can encode. Consistently high IoU across all settings demonstrates the effectiveness of using separate Gaussians to model free and occupied space. Changing ρ or switching between the two pooling rules produces only minor fluctuations, indicating that the method is robust to these settings within the tested range.

Visualizations. Figure 3 presents qualitative results together with two reference annotations: (i) the ground truth restricted to the ego vehicle's field of view (Ego GT) and (ii) the combined ground truth from all connected vehicles (Collaborative GT). After exchanging Gaussian primitives, our model reconstructs more complete instances of vehicles, road surfaces, terrain, walls, and traffic signs than those visible in Ego GT, filling regions that were initially occluded. In some scenarios our prediction is even smoother and more

continuous than Collaborative GT. Compared with the *zero-shot* variant, whose direct stacking of Gaussians produces cluttered occupancy maps (bottom row), the proposed cross-agent Gaussian fusion outputs clean and coherent representations of the scene.

Conclusion

In this work, we demonstrate that explicit 3D Gaussian primitives serve as a compact, interpretable medium for vision-only cooperative 3D semantic occupancy prediction. Our pipeline aligns neighbor Gaussians via rigid transform and region of interest filtering, then refines the ego set with a lightweight, neighborhood-based fusion module. Experiments on Semantic-OPV2V validate that both naive and learned fusion variants achieve notable IoU and mIoU enhancements over planar-feature methods.

References

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4413–4421.

Cao, A.-Q.; and De Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.

Cao, Y.; Jv, Y.; and Xu, D. 2024. 3dgs-det: Empower 3d gaussian splatting with boundary guidance and box-focused sampling for 3d object detection. *arXiv preprint arXiv:2410.01647*.

Chen, X.; Lin, K.-Y.; Qian, C.; Zeng, G.; and Li, H. 2020. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4193–4202.

Ding, F.; Wen, X.; Zhu, Y.; Li, Y.; and Lu, C. X. 2024. Radarocc: Robust 3d occupancy prediction with 4d imaging radar. *Advances in Neural Information Processing Systems*, 37: 101589–101617.

Ding, Z.; Fu, J.; Liu, S.; Li, H.; Chen, S.; Li, H.; Zhang, S.; and Zhou, X. 2025. Point cluster: A compact message unit for communication-efficient collaborative perception. In *The Thirteenth International Conference on Learning Representations*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

He, Y.; Ma, L.; Jiang, Z.; Tang, Y.; and Xing, G. 2021. VI-eye: semantic-based 3D point cloud registration for infrastructure-assisted autonomous driving. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 573–586.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35: 4874–4886.

Hu, Y.; Lu, Y.; Xu, R.; Xie, W.; Chen, S.; and Wang, Y. 2023. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9243–9252.

Huang, T.; Liu, J.; Zhou, X.; Nguyen, D. C.; Azghadi, M. R.; Xia, Y.; Han, Q.-L.; and Sun, S. 2023a. V2X cooperative perception for autonomous driving: Recent advances and challenges. *arXiv preprint arXiv:2310.03525*.

Huang, Y.; Thammatadatrakoon, A.; Zheng, W.; Zhang, Y.; Du, D.; and Lu, J. 2024a. Probabilistic Gaussian Superposition for Efficient 3D Occupancy Prediction. *arXiv preprint arXiv:2412.04384*.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023b. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9223–9232.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2024b. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, 376–393. Springer.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3351–3359.

Li, Y.; Li, S.; Liu, X.; Gong, M.; Li, K.; Chen, N.; Wang, Z.; Li, Z.; Jiang, T.; Yu, F.; et al. 2024. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 13333–13340. IEEE.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34: 29541–29552.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9087–9098.

Liu, Y.-C.; Tian, J.; Ma, C.-Y.; Glaser, N.; Kuo, C.-W.; and Kira, Z. 2020. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 6876–6883. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Roldao, L.; De Charette, R.; and Verroust-Blondet, A. 2022. 3D semantic scene completion: A survey. *International Journal of Computer Vision*, 130(8): 1978–2005.

Shi, S.; Cui, J.; Jiang, Z.; Yan, Z.; Xing, G.; Niu, J.; and Ouyang, Z. 2022. VIPS: Real-time perception fusion for infrastructure-assisted autonomous driving. In *Proceedings of the 28th annual international conference on mobile computing and networking*, 133–146.

Song, R.; Liang, C.; Cao, H.; Yan, Z.; Zimmer, W.; Gross, M.; Festag, A.; and Knoll, A. 2024. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17996–18006.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1746–1754.

Tesla. 2022. Tesla AI Day 2022. https://www.youtube.com/watch?v=ODSJsviD_SU. YouTube video, accessed 2025-07-27.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36: 64318–64330.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European conference on computer vision*, 605–621. Springer.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21729–21740.

Xu, R.; Chen, C.-J.; Tu, Z.; and Yang, M.-H. 2024. V2x-vitv2: Improved vision transformers for vehicle-to-everything cooperative perception. *IEEE transactions on pattern analysis and machine intelligence*.

Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; and Ma, J. 2022a. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*.

Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; et al. 2023. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13712–13722.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022b. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, 107–124. Springer.

Yan, H.; Zheng, Y.; and Duan, Y. 2024. Gaussian-det: Learning closed-surface gaussians for 3d object detection. *arXiv preprint arXiv:2410.01404*.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2023. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36: 25151–25164.

Zhang, J.; Yang, K.; Wang, Y.; Wang, H.; Sun, P.; and Song, L. 2024. Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12575–12584.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9433–9443.

Zhao, L.; Wei, S.; Hays, J.; and Gan, L. 2025. GaussianFormer3D: Multi-Modal Gaussian-based Semantic Occupancy Prediction with 3D Deformable Attention. *arXiv preprint arXiv:2505.10685*.

Zhu, H.; Wang, Y.; Kong, Q.; Wei, Y.; Xia, X.; Deng, B.; Xiong, R.; and Wang, Y. 2024. OTVIC: A Dataset with Online Transmission for Vehicle-to-Infrastructure Cooperative 3D Object Detection. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10732–10739. IEEE.

Zimmer, W.; Wardana, G. A.; Sritharan, S.; Zhou, X.; Song, R.; and Knoll, A. C. 2024. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22668–22677.