Collaborative Semantic Occupancy Prediction with Hybrid Feature Fusion in Connected Automated Vehicles

Rui Song[1,2 *], Chenwei Liang[1], Hu Cao[2], Zhiran Yan[3], Walter Zimmer[2],
Markus Gross[1], Andreas Festag[1,3], Alois Knoll[2]

[1]Fraunhofer IVI [2]Technical University of Munich [3]Technische Hochschule Ingolstadt

https://rruisong.github.io/publications/CoHFF

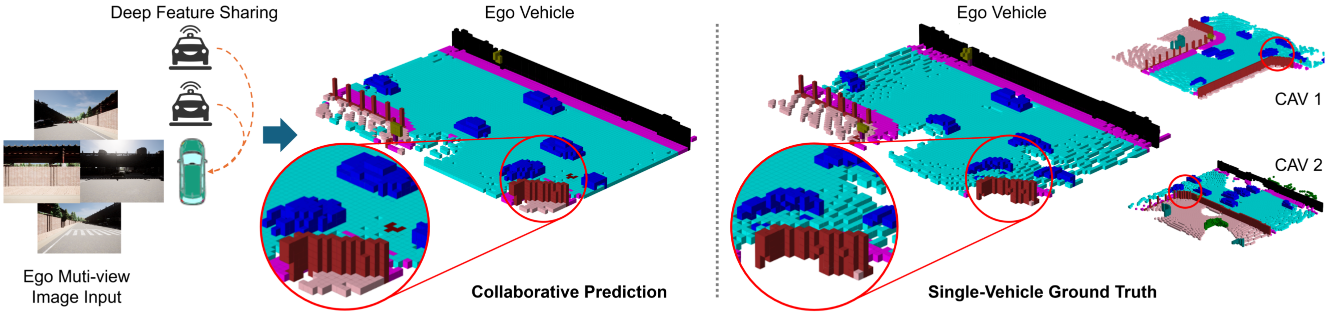

Figure 1. Collaborative semantic occupancy prediction leverages the power of collaboration in multi-agent systems for 3D occupancy prediction and semantic segmentation. This approach enables a deeper understanding of the 3D road environment by sharing latent features among connected automated vehicles (CAVs), surpassing the ground truth captured by a multi-camera system in the ego vehicle.

Abstract

Collaborative perception in automated vehicles leverages the exchange of information between agents, aiming to elevate perception results. Previous camera-based collaborative 3D perception methods typically employ 3D bounding boxes or bird's eye views as representations of the environment. However, these approaches fall short in offering a comprehensive 3D environmental prediction. To bridge this gap, we introduce the first method for collaborative 3D semantic occupancy prediction. Particularly, it improves local 3D semantic occupancy predictions by hybrid fusion of (i) semantic and occupancy task features, and (ii) compressed orthogonal attention features shared between vehicles. Additionally, due to the lack of a collaborative perception dataset designed for semantic occupancy prediction, we augment a current collaborative perception dataset to include 3D collaborative semantic occupancy labels for a more robust evaluation. The experimental findings highlight that: (i) our collaborative semantic occupancy predictions excel above the results from single vehicles by over 30%, and (ii) models anchored on semantic occupancy outpace state-of-the-art collaborative 3D detection techniques in subsequent perception applications, showcasing enhanced accuracy and enriched semantic-awareness in road environments.

1. Introduction

Collaborative perception, also known as cooperative perception, significantly improves the accuracy and completeness of each connected and automated vehicle's (CAV) sensing capabilities by integrating multiple viewpoints, surpassing the limitations of single-vehicle systems [11, 12, 15, 16, 25, 26, 35, 42, 45, 50]. This approach enables CAVs to achieve comparable or superior perception abilities, even with cost-effective sensors. Notably, recent research in [12] suggests that camera-based systems may outperform LiDAR in 3D perception through collaboration in Vehicle-to-Everything (V2X) communication networks. Previous studies in camera-based collaborative perception typically processed inputs from various CAVs into simplified formats such as 3D bounding boxes or Bird's Eye View (BEV) segmentation. While efficient, these methods tend to miss important 3D semantic details, which are indispens-

*Corresponding author, email address: rui.song@ivi.fraunhofer.de able for holistic scene understanding and reliable execution of downstream applications.

Lately, camera-based 3D semantic occupancy prediction, also known as semantic scene completion [31], has become a pioneering method in 3D perception [2, 5, 7, 13, 14, 19, 23, 29, 30, 32, 34, 37–39, 46, 51, 52]. This approach uses RGB camera data to predict the semantic occupancy status of voxels in 3D space, involving both the determination of voxel occupancy and semantic classes of occupied voxels. This research enhances single CAVs' environmental understanding, improving decision-making in downstream applications for automated vehicles. However, this task based on RGB imagery through collaborative methods has not been explored.

To bridge this gap, we delve into the feasibility of 3D semantic occupancy prediction in the context of collaborative perception, as shown in Fig. 1, and introduce the Collaborative Hybrid Feature Fusion (CoHFF) Framework. Our approach involves separate pre-training for the dual subtasks of predicting both semantics and occupancy. We then extract the high-dimensional features from these pretrained models for dual fusion processes: inter-CAV semantic information fusion via V2X Feature Fusion, and intra-CAV fusion of semantic information with occupancy status through task feature fusion. This fusion yields a comprehensive decoding of each voxel's occupancy and semantic details in 3D space.

In order to evaluate the performance of our framework, we extend the existing collaborative perception dataset OPV2V [41]. By reproducing OPV2V scenarios in the CARLA simulator, we collect comprehensive 3D voxel groundtruth with semantic labels across 12 categories. Our experiments show, that for the task of semantic occupancy prediction, a collaborative approach significantly outperforms single-vehicle performance in most categories, as intuitively expected. We also validate the effectiveness of task feature fusion: our findings show that the task fusion, by incorporating features as prior knowledge of each other, enhances subtask performance beyond what separately trained models achieved. Additionally, training tasks independently result in more task-specific features and thus can be easier to compress. Our experiments prove that we achieve more complex 3D perception with a communication volume comparable to existing methods.

Contributions To summarize, our main contributions are threefold:

- We introduce the first camera-based framework for collaborative semantic occupancy prediction, enabling more precise and comprehensive 3D semantic occupancy segmentation than single-vehicle systems through feature sharing in V2X communication networks. The performance can be enhanced by over 30% via collaboration.
- We propose the hybrid feature fusion approach, which not only facilitates efficient collaboration among CAVs, but also markedly enhances the performance over models pre-trained solely for occupancy prediction or semantic voxel segmentation.
- We enrich the collaborative perception dataset OPV2V [41] with voxel ground truth containing 12 categories semantic, bolstering the framework evaluation. Our method, CoHFF, achieves comparable results to current leading methods in subsequent 3D perception applications, and additionally offers more semantic details in road environment.

2. Related work

2.1. Collaborative perception

In intelligent transportation systems, collaborative perception empowers CAVs to attain a more accurate and holistic understanding of the road environment via V2X communication and data fusion. Typically, data fusion in collaborative perception falls into three categories: early, middle, and late fusion. Given the bandwidth limitations of V2X networks, the prevalent approach is middle fusion, where deep latent space features are exchanged [11, 12, 15, 16, 25, 26, 35, 42, 45, 50]. The advantage of middle fusion lies in its ability to convey critical information beyond mere object-level details, bypassing the need to share raw data. The development of datasets specifically designed for collaborative perception [8, 11, 17, 27, 44, 48, 49, 55] has led to remarkable progress in learning-based approaches in recent years. However, these datasets fall short in offering ground truth data for 3D semantic occupancy, which motivates us to extend the dataset in this work, aiming to access the performance of collaborative semantic occupancy prediction.

Collaborative Camera 3D Perception. Compared to LiDAR-driven collaborative perception [44], camera-based methods are often more challenging, due to the absence of explicit depth information in RGB data. However, given the lower price and smaller weight of cameras, they inherently have a higher potential for large-scale deployment. Previous work in [40] and [12] has validated that, with collaboration, camera-based 3D perception can match or even outperform LiDAR performance. Given that current research on camera-based collaborative perception either focuses on 3D bounding box detection and BEV semantic segmentation, there remains a research gap in semantic occupancy prediction. Hence, in this study, our aim is to pioneer and explore the topic of collaborative occupancy segmentation.

2.2. Camera-based semantic occupancy prediction

Occupancy segmentation, which segments a voxel-based 3D environment model [28, 53], has achieved notable success in the realm of autonomous driving. Original occu-

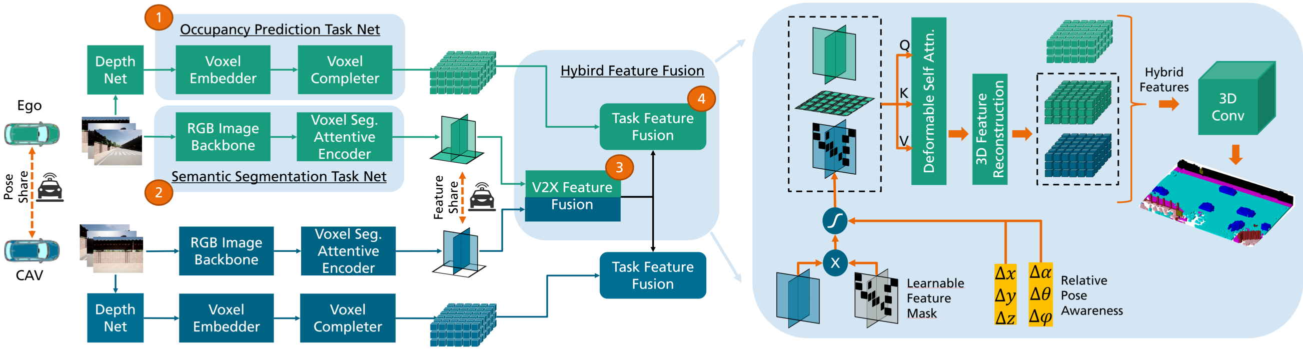

Figure 2. The CoHFF Framework consists of four key modules: (1) Occupancy Prediction Task Net, for occupancy feature extraction; (2) Semantic Segmentation Task Net, creating semantic plane-based embeddings; (3) V2X Feature Fusion, merging CAV features via deformable self-attention; and (4) Task Feature Fusion, uniting all task features to enhance semantic occupancy prediction.

pancy segmentation methods lean heavily on LiDAR, since its point cloud inherits 3D information, aligning naturally with voxel-based environmental models. The recent work proposed in [15] explored the collaborative semantic occupancy prediction based on LiDAR. However, with cameras offering richer environmental details, camera-driven 3D occupancy segmentation is gradually emerging as a novel domain. Recent work in the past year, e.g. [2, 5, 7, 9, 13, 14, 19, 20, 23, 29, 30, 32, 34, 37–39, 51, 52] have also delved into methods for achieving semantic occupancy prediction based on RGB data, yielding promising performance, but only for single vehicle perception.

Furthermore, the datasets for the vision-based 3D Semantic Occupancy Prediction, e.g. Semantic-KITTI [1], SSC-Benchmark [18], OpenOccupancy [36], and Occ3D [33] have been developed specifically for camera-based 3D occupancy segmentation tasks, thus offering resources for continued research. However, those datasets do not support collaborative perception in multi-agent scenarios. Generally, agents sharing different perspective information through collaboration can further enhance voxel-based occupancy segmentation. Due to semantic occupancy prediction offering a more nuanced 3D environmental understanding than collaborative 3D perception methods focused on bounding boxes or BEV perception, it likely requires the exchange of more complex, higher-dimensional features. Determining the most effective information for communication to facilitate the transmission of denser, more informative data stands as a significant challenge.

2.3. Plane-based features

TPVFormer [13] decomposes features for occupancy segmentation into a 3D space. [6] introduced a K-Planes decomposition technique designed to reconstruct static 3D scenes and dynamic 4D videos. Building on the foundations laid by [6], and drawing inspiration from [13], we consider to project semantically relevant information onto orthogonal planes, facilitating information sharing through more streamlined communication. By sharing these plane-based features, we establish the foundational structure of our approach.

3. Methodology

Our CoHFF framework consists of four key modules, namely occupancy prediction Task Net, Semantic Segmentation Task Net, V2X Feature Fusion and Task Feature Fusion, as shown in Fig. 2. It achieves camera-based collaborative semantic occupancy prediction by sharing plane-based semantic features via V2X communication.

3.1. Problem formulation

Given a network of CAVs, defined by a global communication network represented as an undirect graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. For each CAV i, the set of connected CAVs, is denoted by $\mathcal{N}_i = \{j | (i, j) \in \mathcal{E}\}$, where $\mathcal{E}$ is the existing communication links between two CAVs, and j denotes the index of the CAVs connecting to i. We consider the input data in RGB format, and denote $\mathcal{I}_i$ as the image data for a CAV i. The environment model is represented as a 3D voxel grid in one hot embedding $\mathbf{V} \in \mathbb{R}^{X \times Y \times Z \times C}$, where X, Y and Z are voxel grid dimensions. For each CAV i, $\mathbf{V}_i$ represents the predicted occupancy of voxels, while $\mathbf{V}_i^{(0)}$ represents the ground truth of these voxels. The objective of collaborative semantic occupancy prediction, as aligned with the optimization problem in [11, 12], is defined as follows:

$$\max_{\theta, M} \sum_i g(\Phi_\theta(\mathcal{I}_i, \{\mathcal{M}_{i\to j} | j \in \mathcal{N}_i\}), \mathbf{V}_i^{(0)}),$$
$$s.t. \sum_i |\{\mathcal{M}_{i\to j} | j \in \mathcal{N}_i\}| \leq B, \quad (1)$$

where $g(\cdot)$ is the perception metric for optimization. Φ represents the model parametrized by θ, and $\mathcal{M}_{i\to j}$ denotes the message transmitted from CAV i to CAV j. The size of these messages is constrained by a communication budget upper bound $B \in \mathbb{R}^{+}$.

Considering the communication upper bound, instead of directly sending high-dimensional voxel-sized features $\mathcal{F}^{V}$, we opt to transmit features $\mathcal{F}^{\mathbf{P}}$ from orthogonal planes. This approach reduces the messages from $\mathcal{M}^{\mathcal{F}^{V}} \in \mathbb{R}^{X \times Y \times Z \times F}$ to $\mathcal{M}^{\mathbf{P}^{xz}} \in \mathbb{R}^{X \times Z \times F}$ and $\mathcal{M}^{\mathbf{P}^{yz}} \in \mathbb{R}^{Y \times Z \times F}$, where $\mathbf{P}^{xz}$ and $\mathbf{P}^{yz}$ denote the features projected on the xz- and yz-planes respectively. F represents the length of a single feature vector. For instance, in a voxel space of $100 \times 100 \times 8$ with a feature dimension of 128, transmitting orthogonal plane features can reduce communication volume by $50\times$, from 39.05 MB to 0.78 MB, which is comparable to existing collaborative perception methods, yet it offers more extensive and detailed semantic 3D scene information. Based on these considerations, we introduce our framework in the following section.

3.2. Framework design

We divide our method into two distinct pre-communication tasks: 3D occupancy prediction and semantic voxel segmentation. We believe that occupancy features enhance the semantic segmentation performance by providing geometry insight of distinct object classes. Meanwhile, semantic information can suggest changes of a voxel occupancy. Based on this interplay, our approach initially focuses on independent pre-training for each task. Then we fuse the features from both tasks to learn a combined semantic occupancy predictor that yields better performance for each individual task. This assumption is experimentally validated by the ablation study in Tab. 3. Consequently, our framework comprises two specialized pre-trained networks: an occupancy prediction task network and a semantic segmentation task network, as shown in Fig. 2.

Occupancy prediction task network. The occupancy prediction necessitates the conversion of 2D image data into a 3D occupancy grid. We first use an off-the-shelf depth prediction network $\Phi^{depth}(\cdot)$ to determine the depth of each pixel. Following the work in [11, 12], we employ CaDNN [3] for depth estimation. This depth data is then embedded into voxel space through a 3D Emedder, resulting in a preliminary voxel representation. This voxel-based road environment is further completed by a 3D occupancy encoder $\Phi^{occ}(\cdot)$. Finally, the occupancy task features $\mathbf{F}^{occ} \in \mathbb{R}^{X \times Y \times Z \times F}$ is extracted for task fusion.

Semantic segmentation task network. In the segmentation network, we process RGB data to generate feature maps $\mathbf{F}^{seg}$ using $\Phi^{img}(\cdot)$, which are then subjected to deformable cross-attention [54] to facilitate mapping onto a 3D semantic segmentation space. Drawing inspiration from K-Planes [6] and TPVformer [13], we project these features onto three spatially orthogonal planes $\mathcal{P} = \{\mathbf{P}^{xy}, \mathbf{P}^{xz}, \mathbf{P}^{yz}\}$. Among these dense and informative 3D feature representations, two are transmitted via V2X messages, i.e. $\mathcal{M} = \{\mathcal{M}^{\mathbf{P}^{xz}}, \mathcal{M}^{\mathbf{P}^{yz}}\}$. The reason behind not sending the $\mathbf{P}^{xy}$ plane, is that the we use the $\mathbf{P}^{xy}$ of the ego vehicle for reconstructing the 3D features, which facilitates the alignment of the feature space with the detection range of interest of ego vehicle.

Both networks generate high-dimensional features that are fed into a hybrid feature fusion network, thereby forming the core of CoHFF for semantic occupancy prediction.

3.3. Hybrid feature fusion

V2X Feature Fusion. Given one CAV j communicating to the ego vehicle i, the features of the CAV condensed by the segmentation network can contain overlapping information, particularly regarding semantics in proximity to the ego vehicle, which the ego vehicle itself can accurately predict. We implement a masking technique to selectively filter these plane-based features of the CAV, before they are communicated to the ego vehicle. By adjusting a sparsification rate hyperparameter, we reduce the volume of the CAV´s plane-based features shared during collaboration, in line with the communication budget. The compressed message $\bar{\mathcal{M}} = \{\bar{\mathcal{M}}^{\mathbf{P}^{xz}}, \bar{\mathcal{M}}^{\mathbf{P}^{yz}}\}$ can be acquired as follows:

$$\bar{\mathbf{P}}_j^{xz}, \bar{\mathbf{P}}_j^{yz} \leftarrow \mathbf{P}_j^{xz} \odot \mathbf{H}_j^{xz}, \mathbf{P}_j^{yz} \odot \mathbf{H}_j^{yz}, \tag{2}$$

where $\mathbf{H}_j^{xz}$ and $\mathbf{H}_j^{yz}$ represent the learnable feature masks for features on x-z and y-z planes.

Additionally, we ensure relative pose awareness between the ego vehicle and other CAVs. Specifically, we feed the filtered plane features and the relative pose information into an MLP network combined with a Sigmoid function, in line with the methodology proposed in [24].

We now attend these pose-aware filtered plane features from the CAV ($\bar{\mathbf{P}}_j^{xz}$, $\bar{\mathbf{P}}_j^{yz}$) over the three plane features of the ego vehicle ($\mathbf{P}_i^{xy}$, $\mathbf{P}_i^{xz}$, $\mathbf{P}_i^{yz}$). In particular, we use deformable self-attention to update the all five feature planes. The fusion and updating of these planes are accomplished by plane self-attention (PSA), as follows:

$$PSA(\mathbf{p}) = DA(\mathbf{p}, \mathcal{R}, \{\mathbf{P}_i, \bar{\mathbf{P}}_j^{xz}, \bar{\mathbf{P}}_j^{yz} | j \in \mathcal{N}_i\}), \tag{3}$$

where $DA(\cdot)$ is deformable self-attention, $\mathbf{p} \in \mathbb{R}^{F}$ is a query and $\mathcal{R}$ is a set of reference points, as described in [54]. $\mathbf{P}_i$ denotes all the three planes in ego vehicle.

The updated 2D plane features are used in the next step to reconstruct 3D semantic segmentation features $\mathbf{F}^{seg}$. The semantic segmentation feature $\mathbf{f}^{seg}_{x,y,z}$ at a specific Voxel location x, y, z can be reconstructed as follows:

$$\mathbf{f}^{seg}_{x,y,z} = \mathbf{p}^{xy}_{i,z} + \bar{\mathbf{p}}^{xz}_{j,y} + \bar{\mathbf{p}}^{yz}_{j,x} \in \mathbb{R}^{F}, \forall j \in \mathcal{N}_i, \tag{4}$$

where $\bar{\mathbf{p}}^{xz}_{j,y}$ and $\bar{\mathbf{p}}^{yz}_{j,x}$ is plane features from CAV j, and $\mathbf{p}^{xy}_{i,z}$ is the plane (BEV) features from ego vehicle. This idea of

Algorithm 1 : CoHFF framework for collaborative semantic occupancy prediction.

```
for each CAV $i$ in parallel do
    $\mathbf{F}_i^{occ} \leftarrow \Phi^{occ}(Proj(\Phi^{depth}(\mathcal{I}_i), \mathcal{I}_i)))$
    $\mathbf{F}_i^{img} \leftarrow \Phi^{img}(\mathcal{I}_i)$
    update plane-based features $\mathbf{P}_i^{xz}, \mathbf{P}_i^{yz}, \mathbf{P}_i^{xy}$ using deformable cross- and self-attention [54]
    $\bar{\mathbf{P}}_i^{xz}, \bar{\mathbf{P}}_i^{yz} \leftarrow \mathbf{P}_i^{xz} \odot \mathbf{H}_i^{xz}, \mathbf{P}_i^{yz} \odot \mathbf{H}_i^{yz}$
    $\bar{\mathcal{M}}_i \leftarrow \{\bar{\mathbf{P}}_i^{xz}, \bar{\mathbf{P}}_i^{yz}\}$
    CAV $i$ broadcasts messages $\bar{\mathcal{M}}_i$
    for $j \in \mathcal{N}_i$ do
        CAV $i$ receives messages $\bar{\mathcal{M}}_j$
    end for
    update $\{\mathbf{P}_i, \bar{\mathbf{P}}_j^{xz}, \bar{\mathbf{P}}_j^{yz} | j \in \mathcal{N}_i\}$ using self-attention based on (3)
    reconstruct $F_j^{seg}$ based on (4)                                  ▷ VFF
    $\mathbf{V_i} \leftarrow \Phi^{tff}(\mathbf{F}_i^{occ}, \mathbf{F}_i^{seg}, \{\mathbf{F}_j^{seg} | j \in \mathcal{N}_i\})$      ▷ TFF
end for
```

sum of projected features for 3D reconstruction is originally proposed in [13], with our work adapting it to multi-agent scenarios.

Task Feature Fusion. After retrieving global semantic information as $\mathbf{F}^{seg}$, the final step aims at fusion with features $\mathbf{F}^{occ}$ from the occupancy prediction task. To accomplish this, $\mathbf{F}^{seg}$ and $\mathbf{F}^{occ}$ are concatenated and passed to a 3D depth-wise convolution network [47], in order to produce the final semantic voxel map. This task feature fusion network $\Phi^{tff}(\cdot)$ is implemented as follows:

$$\mathbf{V_i} = \Phi^{tff}(\mathbf{F}_i^{occ}, \mathbf{F}_i^{seg}, \{\mathbf{F}_j^{seg} | j \in \mathcal{N}_i\}) \in \mathbb{R}^{X \times Y \times Z \times C}. \tag{5}$$

The CoHFF pseudocode is given in Algorithm 1.

3.4. Losses

We train the completion network training using focal loss proposed in [21], applying it to a dataset with binary labels $\{0, 1\}$. For both the segmentation network and the hybrid feature fusion network, we employ a weighted cross-entropy loss to train for semantic labels. Notably, in this context, the label for the *empty* is also designated as 0.

4. Dataset

To effectively evaluate collaborative semantic occupancy prediction, a dataset that supports collaborative perception and includes 3D semantic occupancy labels is crucial. Thus, we enhance the OPV2V dataset [41] by integrating 12 different 3D semantic occupancy labels, as shown in Tab. 4 This enhancement is achieved using the high-fidelity CARLA simulator [4] and the OpenCDA autonomous driving simulation framework [43]. We position four semantic LiDARs at the original camera sites to precisely capture the semantic occupancy ground truth within the cameras' FoV. In addition, we associate ground truth data from all CAVs to create a detailed collaborative ground truth for collaborative supervision. Furthermore, to comprehensively capture occluded semantic occupancies for all CAVs, we include a simulation replay in our data collection process, where each CAV is equipped with 18 semantic LiDARs. This strategic configuration is crucial for effectively evaluating completion tasks, as it guarantees extensive data collection, encompassing areas not visible in direct associated FoV. In alignment with the original OPV2V protocol, we replay the simulation and generate a multi-tier ground truth.

Table 1. Comparison 3D object detection with AP[2] of vehicles.

Approach	# Agents	AP@0.5	AP@0.7
DiscoNet (NeurIPS 21)	Up to 7	36.00	12.50
V2X-ViT (ECCV 22)	Up to 7	39.82	16.43
Where2Comm (NeurIPS 22)	Up to 7	47.30	19.30
CoCa3D (CVPR 23)	7[1]	**69.10**	**49.50**
CoHFF	Up to 7	48.51	36.39
CoCa3D-2 (CVPR 23)	2	25.90	12.60
CoHFF	2	**36.63**	**27.95**

[1] CoCa3D is trained on OPV2V+, where extended agents provide more input information for better results.

[2] We calculate the 3D IoU by comparing the predicted voxels with the ground truth voxels for each object, rather than using 3D bounding boxes due to the potential unnecessary occupancy in 3D bounding boxes.

Table 2. Comparison of BEV semantic segmentation with IoU in the class of Vehicle, Road and Others.

Approach	# Agents	Vehicle	Road	Others[1]
CoBEVT (CoRL 22)	2	46.13	52.41	-
CoHFF	2	**47.40**	**63.36**	**40.27**
CoBEVT (CoRL 22)	Up to 7	60.40	**63.00**	-
CoHFF	Up to 7	**64.44**	57.28	**45.89**

[1] It refers to additional object classes identified through semantic segmentation predictions projected onto the BEV plane. These categories include buildings, fences, terrain, poles, vegetation, walls, guard rails, traffic signs, and bridges. The IoU for these objects is calculated and reported as IoU.

5. Experimental evaluation

5.1. Experiment setup

Baselines. Considering the unexplored domain of collaborative occupancy segmentation, we extend the findings from CoHFF to address downstream applications, including BEV perception and 3D detection. In our analysis, we evaluate these outcomes with those from state-of-the-art collaborative perception models that employ multi-view cameras: CoBEVT [40] for BEV perception and CoCa3D [12] for 3D detection. Furthermore, we examine contemporary

Table 3. CoHFF achieves robust IoU and mIoU performance, when the communication volume (CV) is reduced by setting various sparsification rates (Spar. Rate).

Spar. Rate	0.00	0.50	0.80	0.95	0.99
CV (MB) (↓)	16.53	8.27	3.31	0.83	0.17
IoU (↑)	50.46	49.56	49.53	48.52	48.02
mIoU (↑)	34.16	32.97	32.70	30.13	29.48

methods that integrate alternative modalities, particularly those blending LiDAR with camera inputs or relying solely on LiDAR, including DiscoNet [16], V2X-ViT [42] and Where2Comm [11].

Implementation details. Following the previous work for collaborative perception evaluation on the OPV2V dataset used in [11], we utilize a $40 \times 40 \times 3.2$ meter detection area with a grid size of 100 x 100 x 8, resulting in a voxel size of $0.4\ m^3$. We allow CAVs to transmit and share features with a length of 128 for V2X Feature Fusion. Our experiment incorporates the analysis of 12 semantic labels plus an additional *empty* label. We employ CaDNN [3] with 50 depth categories and a single out-of-range category for depth estimation, as well as ResNet101 [10] and FPN [22] as RGB the image backbone. For Voxel completion, we utilize a 3D depth-wise CNN [47] and use deformable attention [54] in hybrid feature fusion.

Evaluation metrics. Following the evaluation of semantic occupancy prediction in previous work, such as [2, 13, 19], we primarily utilize the metric Intersection over Union (IoU) for evaluation. This involves calculating IoU for each individual class and the mean IoU (mIoU) across all classes. Additionally, for evaluations in subsequent applications, we compute the Average Precision (AP) at IoU threshold of 0.5 and 0.7, and BEV 2D IoU to compare with other baselines. Specifically, the AP value is calculated only for voxels labeled as vehicles, and the IoU is determined for each pair of predicted and actual vehicles. For BEV IoU, voxels are projected onto the BEV plane and categorized into the corresponding semantic classes.

5.2. Comparison

Collaborative 3D object detection. First, we compare the performance of CoHFF in 3D detection applications. As shown in Tab. 1, with up to 7 agents' collaborative perception, CoHFF achieves comparable performance to Where2comm at AP@0.5 and obtains an 88.5% improvement at AP@0.7. We believe this is primarily due to semantic occupancy prediction, which makes the perception results closer to the actual observed shapes, rather than inferring a non-existent bounding box in the scenarios. We also observe that CoCa3D, on the OPV2V+ dataset [12], achieves significantly better performance due to receiving more information from CAVs. To compare directly with CoCa3D, we also conduct scenarios where only two agents communicated at a time. We can see that CoHFF has made significant improvements at both AP@0.5 and AP@0.7.

Collaborative BEV segementation. Tab. 2 presents a comparison between CoHFF and CoBEVT in BEV semantic segmentation. Note that errors in height prediction from 3D voxel occupancy mapping to the BEV plane may be overlooked during the projection process. Despite this, CoHFF achieves even better performance in predicting vehicles and roads in BEV compared to CoBEVT. Additionally, CoHFF is capable of detecting a wider range of other semantic categories in 3D occupancy.

5.3. Ablation study

To validate our hypothesis that independently obtained semantic and occupancy feature information can simultaneously strengthen the original semantic and occupancy tasks, we have decomposed the semantic occupancy prediction into two separate tasks. Tab. 4 shows an ablation study by altering the components used. Meanwhile, we also verify the enhancement of collaborative perception over single vehicle perception in terms of semantic occupancy.

CoHFF for occupancy prediction. When focusing solely on binary occupancy predictions (as shown at Occ. Pred. in Tab. 4), we use voxels processed from raw LiDAR point clouds as a reference, and analyze the IoU in different semantic classes based on semantic occupancy in ground truth. It is observed that by utilizing an occupancy prediction task network to process depth predictions, the overall prediction accuracy is enhanced. Additionally, significant improvements in predicting large objects in occupancy results are noted by integrating features from a semantic segmentation task network, leading to an increased overall IoU. However, a concurrent decline in the mIoU is observed alongside the increase in IoU. This phenomenon is attributed to the influence of semantic features, which seem to steer the model towards prioritizing easily detectable categories, potentially at the expense of smaller or less distinct categories. Finally, through collaboration, the overall IoU and mIoU are further strengthened on the basis of task feature fusion.

CoHFF for semantic segmentation. In our semantic segmentation task (as shown at Sem. Seg. in Tab. 4), after integrating features from occupancy prediction, we observe an approximate 2% increase in IoU, but a more substantial over 41% enhancement in mIoU. We attribute this improvement to the features derived from occupancy prediction, which seem to aid the easier detection of smaller-scale objects, thereby refining their semantic predictions. Consistent with the occupancy prediction task, the final collaboration further elevates the results of semantic segmentation.

Table 4. Component ablation study on occupancy prediction (Occ. Pred.), semantic segmentation (Sem. Seg.), and semantic occupancy prediction (Sem. Occ. Pred.) tasks. The components include: Occupancy Prediction Task Net (OPTN), Semantic Segmentation Task Net (SSTN), Task Feature Fusion (TFF) and V2X Feature Fusion (VFF). The gray color in table cells indicates that the corresponding component is not applicable for the task.

Task type	Occ. Pred.				Sem. Seg.			Sem. Occ. Pred.	
OPTN	RL [1]	✓	✓	✓				✓	✓
SSTN					✓	✓	✓	✓	✓
TFF			✓	✓		✓	✓	✓	✓
VFF (Collaboration)				✓			✓		✓
IoU (↑)	49.35	67.22	76.62	**86.37**	41.30	42.11	**51.38**	38.52	**50.46**
mIoU (↑)	57.12	64.01	59.16	**69.15**	21.59	30.51	**35.91**	24.85	**34.16**
Building (5.40%) 🟥	67.50	**68.36**	41.29	48.41	9.65	**27.25**	15.06	21.04	**25.72**
Fence (0.85%) 🟫	59.40	62.05	51.60	**65.01**	11.67	30.29	**30.91**	20.50	**27.83**
Terrain (4.80%) 🟥	43.60	49.78	68.21	**79.81**	51.18	51.41	**61.98**	43.93	**48.30**
Pole (0.39%) 🟨	66.30	**70.67**	62.31	64.12	2.14	36.80	**40.74**	31.66	**42.74**
Road (40.53%) 🟦	51.47	77.78	91.26	**93.00**	56.82	60.02	**64.09**	55.83	**61.77**
Side walk (35.64%) 🟪	45.46	58.46	74.37	**90.53**	25.22	16.87	**36.03**	17.31	**39.62**
Vegetation (1.11%) 🟩	43.61	**44.43**	38.87	41.57	9.12	**22.13**	20.99	14.49	**20.59**
Vehicles (9.14%) 🟦	41.40	63.53	59.52	**76.48**	59.58	69.81	**75.88**	58.55	**63.28**
Wall (2.01%) ⬛	71.51	79.35	49.63	**81.20**	32.55	39.80	**58.49**	33.30	**58.27**
Guard rail (0.04%) 🟪	**49.67**	46.03	41.35	43.33	1.10	**1.95**	1.80	1.54	**1.94**
Traffic signs (0.05%) 🟫	68.98	**69.41**	52.35	62.54	0.00	9.77	**11.69**	0.00	**16.33**
Bridge (0.04%) 🟦	76.53	78.23	79.08	**83.84**	0.00	0.00	**13.30**	0.00	**3.53**

[1] RL (Raw LiDAR) is used as a baseline for the evaluation on the task of occupancy prediction.

Collaboration enhances semantic occupancy prediction. In the final evaluation of our semantic occupancy prediction (see column Sem. Occ. Pred. in Tab. 4), we further demonstrate the benefits brought by collaboration. By collaboration, the IoU for each category is improved. Notably, some previously undetectable, low-prevalence categories such as traffic signs and bridges can be detected after collaboration. Ultimately, there is an approximate **31%** increase in overall IoU and around a **37%** enhancement in mIoU.

5.4. Robustness with low communication budget

In Tab. 3, we increase the sparsification rate to mask the plane-based features transmitted by CAVs, achieving efficient V2X information exchange under a low communication budget. The CoHFF model exhibits stable IoU performance across various levels of sparsification. Even when the communication volume is shrinked by 97 ×, the accuracy only decreases by 5% compared to the original. Meanwhile, the mIoU drops by 15%. Despite this, due to the model's training under collaborative supervision, it still outperforms the non-collaborative approach.

5.5. Visual analysis

Fig. 3 presents visual results from the CoHFF model, which are compared from multiple perspectives with the ground truth data, i.e. the ground truth in the ego vehicle's FoV (Ego GT) and the ground truth across all CAVs FoVs (Collaborative GT). It is evident that, overall, the model accurately predicts voxels in various classes such as roads, sidewalks, traffic signs, walls, and fences. We particularly focus on vehicle predictions, as they are among the most critical categories in road environment perception. For clarity, each vehicle object in the figure is numbered.

Vehicle geometry completion. The CoHFF model predicts more complete vehicle objects than those in the Ego GT, such as vehicles 1, 3, 4, and 7. In some instances, the predictions even surpass the completeness of vehicle shapes found in Collaborative GT.

Occluded vehicle detection. CoHFF successfully predicts vehicles outside of the FoV, such as vehicle 6, by utilizing minimal pixel information. This demonstrates that CoHFF can effectively detect occluded vehicles.

6. Conclusion

In this work, we explore the task of camera-based semantic occupancy prediction through the lens of collaborative perception. We introduce the CoHFF framework, which significantly enhances the perception performance by over 30% through integrating features from different tasks and various CAVs. Since currently no dataset specifically de-

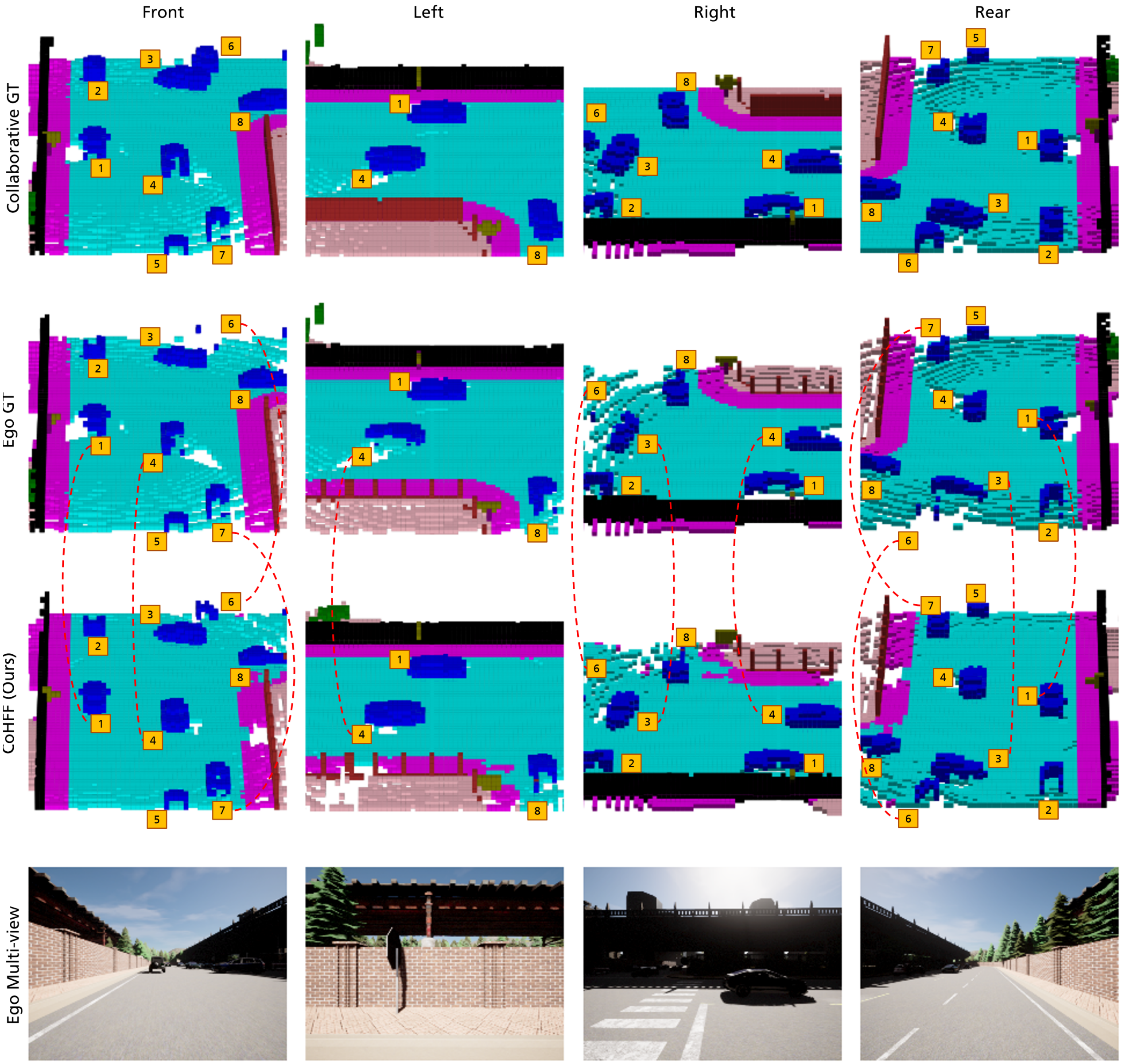

Figure 3. Illustration of collaborative semantic occupancy prediction from multiple perspectives, compared to the ground truth in the ego vehicle's FoV and the collaborative FoV across CAVs. This visualization emphasizes the advanced object detection capabilities in collaborative settings, particularly for objects obscured in the ego vehicle's FoV, such as the vehicle with ID 6.

signed for collaborative semantic occupancy prediction exists, we also extend the OPV2V dataset with 3D semantic occupancy labels. Our experiments validate that collaboration yields better semantic occupancy prediction results than single-vehicle approaches.

Limitation. Although we demonstrate the immense potential of collaboration for semantic occupancy prediction using simulation data, its performance with real-world data remains to be verified. The collection and development of a specialized dataset, repleted with semantic occupancy labels and derived from multi-agent perception scenarios in real-world settings, are highly anticipated.

7. Acknowledgements

This work was supported by the German Federal Ministry for Digital and Transport (BMVI) in the project ”5GoIng – 5G Innovation Concept Ingolstadt”.

References

[1] J. Behley et al. Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset. *The International Journal on Robotics Research*, 40(8-9):959–967, 2021. DOI: 10.1177/02783649211006735. 3

[2] Anh-Quan Cao and Raoul de Charette. MonoScene: Monocular 3D semantic scene completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3991–4001. IEEE, 2022. DOI: 10.1109/CVPR52688.2022.00396. 2, 3, 6

[3] Cody Ceading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution network for monocular 3D object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8551–8560, 2021. DOI: 10.1109/CVPR46437.2021.00845. 4, 6

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 5

[5] Shaoheng Fang, Zi Wang, Yiqi Zhong, Junhao Ge, and Siheng Chen. TBP-Former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1368–1378. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.00138. 2, 3

[6] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12488. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.01201. 3, 4

[7] Aditya Nalgunda Ganesh, Dhruval Pobbathi Badrinath, Harshith Mohan Kumar, Priya SS, and Surabhi Narayan. OCTraN: 3D occupancy convolutional transformer network in unstructured traffic scenarios. *arXiv preprint arXiv:2307.10934*, 2023. 2, 3

[8] R Hao et al. Rcooper: A real-world large-scale dataset for roadside cooperative perception. *arXiv preprint arXiv:2403.10145*, 2024. 2

[9] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4C: Self-supervised semantic scene completion with neural fields. *arXiv preprint arXiv:2310.07522*, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. DOI: 10.1109/CVPR.2016.90. 6

[11] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 4874–4886, 2022. 1, 2, 3, 4, 6

[12] Yue Hu et al. Collaboration helps camera overtake LiDAR in 3D detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.00892. 1, 2, 3, 4, 5, 6

[13] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3D semantic occupancy prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9223–9232. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.00890. 2, 3, 4, 5, 6

[14] Haoyi Jiang et al. Symphonize 3D semantic scene completion with contextual instance queries. *arXiv preprint arXiv:2306.15670*, 2023. 2, 3

[15] Yiming Li, Juexiao Zhang, Dekun Ma, Yue Wang, and Chen Feng. Multi-robot scene completion: Towards task-agnostic collaborative perception. In *Conference on Robot Learning*, pages 2062–2072. PMLR, 2023. 1, 2, 3

[16] Yiming Li et al. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:29541–29552, 2021. 1, 2, 6

[17] Yiming Li et al. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. DOI: 10.1109/LRA.2022.3192802. 2

[18] Yiming Li et al. SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving. *arXiv preprint arXiv:2306.09001*, 2023. 3

[19] Yiming Li et al. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.00877. 2, 3, 6

[20] Zhiqi Li et al. FB-OCC: 3D occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 3

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. DOI: 10.1109/TPAMI.2018.2858826. 5

[22] Tsung-Yi Lin et al. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. DOI: 10.1109/CVPR.2017.106. 6

[23] Jihao Liu et al. Towards better 3D knowledge transfer via masked image modeling for multi-view 3D understanding. *arXiv preprint arXiv:2303.11325*, 2023. 2, 3

[24] Yingfei Liu et al. Petrv2: A unified framework for 3D perception from multi-camera images. In *2023 IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3262–3272, 2023. 4

[25] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *2020 IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition (CVPR), pages 4106–4115. IEEE, 2020. DOI: 10.1109/CVPR42600.2020.00416. 1, 2

[26] Yen-Cheng Liu et al. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883. IEEE, 2020. DOI: 10.1109/ICRA40945.2020.9197364. 1, 2

[27] C. Ma et al. Holovic: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative. *arXiv preprint arXiv:2403.02640*, 2024. 2

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. DOI: 10.1109/CVPR.2019.00459. 2

[29] Ruihang Miao et al. OccDepth: A depth-aware method for 3D semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023. 2, 3

[30] Chen Min et al. Occ-BEV: Multi-camera unified pre-training via 3D scene reconstruction. *arXiv preprint arXiv:2305.18829*, 2023. 2, 3

[31] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3D semantic scene completion: A survey. *International Journal of Computer Vision*, 130(8):1978–2005, 2022. 2

[32] Zhiyu Tan et al. Ovo: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023. 2, 3

[33] Xiaoyu Tian et al. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 3

[34] Wenwen Tong et al. Scene as occupancy. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8406–8415, 2023. 2, 3

[35] Tsun-Hsuan Wang et al. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In *2020 European Conference on Computer Vision (ECCV)*, pages 605–621, Glasgow, UK, 2020. Springer. DOI: 10.1007/978-3-030-58536-5_36. 1, 2

[36] Xiaofeng Wang et al. OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17850–17859, 2023. 3

[37] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified occupancy representation for camera-based 3D panoptic segmentation. *arXiv preprint arXiv:2306.10013*, 2023. 2, 3

[38] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. PET-NeuS: Positional encoding tri-planes for neural surfaces. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12598–12607. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.01212.

[39] Yi Wei et al. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21729–21740, 2023. 2, 3

[40] Runsheng Xu et al. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022. 2, 5

[41] Runsheng Xu et al. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. DOI: 10.1109/ICRA46639.2022.9812038. 2, 5

[42] Runsheng Xu et al. V2X-VIT: Vehicle-to-everything cooperative perception with vision transformer. In *2022 European Conference on Computer Vision (ECCV)*, pages 107–124. Springer, 2022. DOI: 10.1007/978-3-031-19842-7_7. 1, 2, 6

[43] Runsheng Xu et al. The OpenCDA open-source ecosystem for cooperative driving automation research. *IEEE Transactions on Intelligent Vehicles*, 8(4):2698–2711, 2023. DOI: 10.1109/TIV.2023.3244948. 5

[44] Runsheng Xu et al. V2V4Real: a real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13712–13722. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.01318. 2

[45] Kun Yang et al. Spatio-temporal domain awareness for multi-agent collaborative perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23383–23392, 2023. 1, 2

[46] Jiawei Yao et al. NDC-Scene: Boost monocular 3D semantic scene completion in normalized device coordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9455–9465, 2023. 2

[47] Rongtian Ye, Fangyu Liu, and Liqiang Zhang. 3D depthwise convolution: Reducing model parameters in 3D vision tasks. In *32nd Canadian Conference on Artificial Intelligence (Canadian AI), Proc. 32*, pages 186–199. Springer, 2019. DOI: 10.1007/978-3-030-18305-9_15. 5, 6

[48] Haibao Yu et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21361–21370. IEEE, 2022. DOI: 10.1109/CVPR52688.2022.02067. 2

[49] Haibao Yu et al. V2X-Seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5486–5495. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.00531. 2

[50] Haibao Yu et al. Vehicle-infrastructure cooperative 3D object detection via feature flow prediction. *arXiv preprint arXiv:2303.10552*, 2023. 1, 2

[51] Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9433–9443, 2023. 2, 3

[52] Zaibin Zhang, Lijun Wang, Yifan Wang, and Huchuan Lu. BEV-IO: Enhancing bird's-eye-view 3D detection with instance occupancy. *arXiv preprint arXiv:2305.16829*, 2023. 2, 3

[53] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3D object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2018. DOI: 10.1109/CVPR.2018.00472. 2

[54] Xizhou Zhu et al. Deformable DETR: Deformable transformers for end-to-end object detection. In *2021 International Conference on Learning Representations (ICLR)*, 2021. 4, 5, 6

[55] W. Zimmer et al. TUMTraf V2X cooperative perception dataset. *arXiv preprint arXiv:2403.01316*, 2024. 2