2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

A Synthetic Benchmark for Collaborative 3D Semantic Occupancy Prediction in V2X-Enabled Autonomous Driving

Hanlin Wu, Pengfei Lin*, Ehsan Javanmardi, Naren Bao, Bo Qian, Hao Si, Manabu Tsukada

Abstract— **3D semantic occupancy prediction is an emerging perception paradigm in autonomous driving, providing a voxel-level representation of both geometric details and semantic categories. However, despite its fine-grained scene understanding, its effectiveness is inherently constrained in single-vehicle setups by occlusions, restricted sensor range, and narrow viewpoints. To address these limitations, collaborative perception enables the exchange of complementary information, thereby enhancing the completeness and accuracy of predictions. Despite its potential, research on collaborative 3D semantic occupancy prediction is hindered by the lack of dedicated datasets. To bridge this gap, we design a high-resolution semantic voxel sensor in CARLA to produce dense and comprehensive annotations for V2X scenarios. We further develop a baseline model that performs inter-agent feature fusion via spatial alignment and attention aggregation. In addition, we establish benchmarks with varying prediction ranges designed to systematically assess the impact of spatial extent on collaborative prediction. Experimental results demonstrate the superior performance of our baseline enabled by vehicle collaboration, with increasing gains observed as the prediction range expands. Our codes and data are available at https://github.com/tlab-wide/Co3SOP.**

I. Introduction

Collaborative perception has emerged as a powerful strategy to enhance scene understanding in autonomous driving [1], [2]. By sharing sensory information across multiple agents, it enables vehicles to perceive a broader environment beyond their individual field of view. This is particularly crucial in urban driving scenarios where occlusions and limited sensor range frequently hinder the perception performance of a single vehicle. Prior research [3], [4], [5] has shown that inter-agent communication and feature fusion can significantly improve the accuracy of 3D detection, segmentation, and tracking tasks.

However, most existing collaborative perception methods employ coarse environment representations, such as 3D bounding boxes or bird's-eye-view (BEV) maps. While effective for certain tasks, these representations are insufficient for capturing the fine-grained geometry and semantics necessary for downstream reasoning and planning. Recently, 3D semantic occupancy prediction, also known as Semantic Scene Completion (SSC), has gained attention for its ability to provide detailed 3D semantic and geometric information by utilizing a fine-grain voxel-based representation [6], [7], [8], offering a richer and more detailed understanding of the environment.

Despite its potential, 3D semantic occupancy prediction remains largely unexplored in collaborative settings. This is due in part to the absence of dedicated datasets and benchmarks that provide voxel-level semantic ground-truth annotations under collaborative settings [9], [10]. Existing datasets, such as SemanticKITTI [11] and Occ3D [12], lack support for multi-agent configurations and primarily rely on LiDAR point cloud of single vehicle to generate annotations. However, the inherent sparsity of LiDAR point clouds, particularly at greater distances, coupled with occlusions, sensor noise, and non-uniform point distributions, poses significant challenges to generating accurate and reliable annotations. Moreover, collaborative perception imposes higher demands on annotations, requiring temporally and spatially aligned multi-agent labels to ensure consistent supervision and effective fusion.

To address this gap, we augment an existing multi-agent dataset [13] by replaying it in CARLA with a high-resolution semantic voxel sensor, yielding dense voxel-level annotations. The resulting dataset, which we term Co3SOP, supports the training and evaluation of collaborative 3D semantic occupancy prediction models under simulation scenarios. Additionally, we introduce a baseline model that incorporates inter-agent feature fusion with spatial alignment and sparse attention. Specifically, we apply warping-based spatial alignment to transform neighboring agents' features into the ego frame, followed by a visibility-guided sparse attention mechanism modulated by a learned confidence mask to adaptively weight contributions from each agent. Alongside the dataset and baseline, we establish benchmarks with different prediction ranges, to systematically assess how spatial distance influences the effectiveness of collaboration.

To benchmark the effectiveness of our dataset and proposed baseline model, we conduct extensive experiments on our proposed dataset, evaluating a range of state-of-the-art single-agent models as well as our collaborative baseline across all benchmark splits. The results demonstrate that collaborative perception significantly boosts semantic occupancy prediction performance, with larger collaboration ranges yielding greater improvements. In addition, we further incorporate pose noise simulation to approximate real-world uncertainty in inter-agent communication and sensing. These findings reinforce the potential of voxel-based representations in collaborative 3D scene understanding and establish Co3SOP as a foundation for future research in this direction.

Hanlin Wu, Pengfei Lin, Ehsan Javanmardi, Naren Bao, Bo Qian, Hao Si and Manabu Tsukada are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113-8657 Japan. (e-mail: {hanlinwu, linpengfei0609, ejavanmardi, naren}@g.ecc.u-tokyo.ac.jp, boqian@ieee.org, {si-hao, mtsukada}@g.ecc.u-tokyo.ac.jp)

* Corresponding to linpengfei0609@g.ecc.u-tokyo.ac.jp

TABLE I: **Comparison of the existing 3D Occupancy Prediction Dataset for Autonomous Driving and our proposed Co3SOP**. Real and Sim represent the raw data is collected from real world or simulation platform. Gen means the 3D occupancy annotations are generated from raw data. C, L denote camera and LiDAR.

Dataset	Source	Meta Dataset	Modality	Voxels Size	Resolution	V2X Support
SemanticKITTI	Real+Gen	KITTI	C&L	[256, 256, 32]	[0.2, 0.2, 0.2]	-
KITTI-360	Real+Gen	KITTI	C&L	[256, 256, 32]	[0.2, 0.2, 0.2]	-
Occ3D-nuScenes	Real+Gen	nuScenes	C&L	[200, 200, 16]	[0.4, 0.4, 0.4]	-
Occ3D-Waymo	Real+Gen	Waymo	C&L	[3200, 3200, 128]	[0.05, 0.05, 0.05]	-
SSCBench	Real+Gen	nuScenes&Waymo &KITTI-360	C&L	[256, 256, 32]	[0.2, 0.2, 0.2]	-
V2VSSC	Sim+Gen	OPV2V	C&L	[128, 128, 20]	[0.78, 0.78, 0.4]	V2V
Co3SOP (Ours)	Sim	OPV2V	C&L	[1000, 1000, 70]	[0.1, 0.1, 0.1]	V2V

II. Related Work

A. 3D Semantic Occupancy Prediction

Methodologies: 3D semantic occupancy prediction aims to provide voxel-level understanding of both geometry and semantics, and has seen rapid development across different sensing modalities. Early approaches such as SSCNet [14] adopt volumetric CNNs on RGB-D inputs. LiDAR-based methods like LMSCNet [15], S3CNet [16], and JS3CNet [17] exploit the structural sparsity and accuracy of 3D point clouds to produce high-quality voxel predictions. Meanwhile, vision-based methods, including MonoScene [18], infer 3D semantics by lifting monocular depth and segmentation into voxel space. To overcome the limitations of single-view inputs, recent works such as VoxFormer [19] and OccFormer [20] leverage multi-camera fusion and transformer-based 2D-to-3D attention mechanisms. Despite steady progress, these methods are restricted to single-agent perception, and thus remain vulnerable to occlusion, field-of-view limitations, and sensor sparsity.

Datasets: To facilitate the development of 3D semantic occupancy prediction, several datasets have been proposed with voxel-level annotations derived primarily from LiDAR scans. SemanticKITTI [11], [21], [22] is a widely adopted benchmark, providing dense semantic labels on voxelized point clouds. KITTI-360 [23] extends this setup with panoramic imagery and 360° LiDAR coverage. Occ3D [12] introduces a voxel annotation pipeline based on Waymo and nuScenes data, offering large-scale semantic occupancy labels. SSCBench [24] further consolidates multiple sources to unify evaluation across diverse urban scenes. While these datasets advance single-agent occupancy prediction, they do not support multi-agent collaboration. V2VSSC [25] is the first to target vehicle-to-vehicle (V2V) collaborative occupancy tasks, yet its annotations remain limited by the sparsity and occlusion inherent in LiDAR data, compromising label density and accuracy. In contrast, our proposed Co3SOP dataset offers simulation-based dense annotations specifically tailored for collaborative settings, as summarized in table I.

B. Collaborative Perception

Methodologies: Collaborative perception enhances environmental understanding by enabling vehicles to exchange complementary information. A variety of methods have been proposed to realize effective multi-agent fusion. V2VNet [26] adopts graph neural networks to aggregate features from spatially distributed agents. V2X-ViT [3] leverages vision transformers to model inter-agent interactions through attention mechanisms. To improve communication efficiency, Where2comm [4] identifies task-relevant regions for selective transmission, while How2comm [5] explores policies for adaptive communication scheduling. While effective for tasks such as object detection and segmentation, these approaches are limited to object-centric outputs and fall short in supporting fine-grain voxel-level semantic understanding.

Datasets: Ego-vehicle datasets such as KITTI [27], nuScenes [28], and Waymo [27] fall short in supporting collaborative perception, as they lack multi-agent scenarios and aligned multi-view annotations. To address these limitations, a range of specialized datasets have emerged [9], [10], [29], [30]. Among simulated datasets, V2X-Sim [31] models V2X interactions in CARLA and supports tasks such as detection, tracking, and BEV segmentation. OPV2V [13], also built in CARLA, emphasizes real-time V2V communication and provides configurable multi-agent scenarios. In real-world settings, V2V4Real [32] focuses on vehicle-to-vehicle collaboration, while DAIR-V2X [33] targets vehicle-to-infrastructure (V2I) scenarios. Additionally, V2X-Seq [34] supports sequential V2X perception by enabling temporal information sharing across agents. However, none of the above datasets provide dense voxel-level semantic annotations required for collaborative 3D semantic occupancy prediction.

III. Co3SOP Dataset

A. Semantic Voxel Annotation Pipeline

We construct the Co3SOP dataset by replaying multi-agent driving scenarios from OPV2V [13] in the CARLA simulator and augmenting them with dense voxel-level semantic occupancy labels. Unlike prior datasets that rely on sparse LiDAR observations, Co3SOP leverages CARLA's high-fidelity simulation environment to generate complete ground truth.

While CARLA offers a variety of sensors, it does not natively support 3D semantic voxel output. To address this

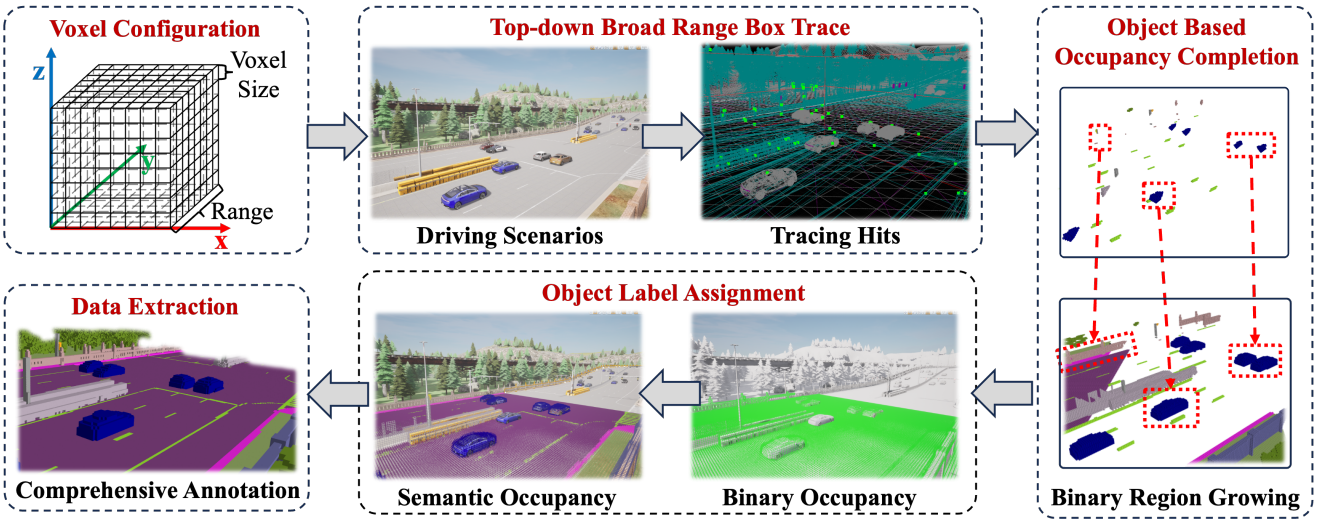

Fig. 1: Illustration of the annotation pipeline for 3D semantic voxel using a custom sensor in Carla, including voxel configuration, top-down broad range box trace, object based occupancy completion and label assignment.

limitation, we develop a custom semantic voxel sensor using Unreal Engine's built-in collision and overlap detection functions. This sensor efficiently retrieves both occupancy and semantic information at the voxel level, enabling fine-grained annotations for collaborative occupancy prediction tasks.

Dividing a large scene into high-resolution voxels and checking each voxel individually, however, can introduce significant computational overhead. For example, a detection range of $100 \times 100 \times 4.8\,m^3$ at $0.1\,m$ resolution results in approximately 48 million detection operations for one frame. To alleviate this, we design a multi-stage voxel annotation pipeline, as illustrated in fig. 1 that leverages scene-level object cues to minimize redundant computations:

Sensor Configuration. The voxel is centered at each ego vehicle with specific sensor parameters, including the voxel range and resolution, which are configured to define the spatial extent and granularity. These settings determine how the scene is divided into 3D voxels.

Top-Down Broad-Range Box Trace. After configuration, a top-down overlap detection is performed to coarsely identify all physical objects intersecting with the voxelized scene. Each object's intersection points (i.e., impact voxels) are recorded and used to initialize a list of seed voxels likely to be occupied. This step refines the focus of annotation to areas likely containing occupied voxels, streamlining further annotation.

Object-Level Occupancy Completion. Then, for each object, a parallel Breadth-First Search (BFS) is initiated from its seed voxels. The BFS selectively propagates to the six axis-aligned neighboring voxels (i.e., $\pm x, \pm y$, and $\pm z$), where each candidate undergoes a collision-aware check for occupancy. Occupied occupied voxels are added to the expansion frontier for subsequent iterations. This localized and iterative process continues until all reachable occupied voxels are exhaustively annotated. The entire design is inherently multi-threaded, with each object assigned to an independent thread, ensuring efficient and scalable occupancy annotation.

Object Label Assignment. After occupancy completion, the semantic category of the intersecting object—retrieved from the CARLA simulation engine—is assigned to each of its occupied voxels, completing the voxel-level annotation process.

This optimized annotation pipeline ensures fine-grained spatial annotations by focusing computation on relevant voxel regions, thereby avoiding redundant processing. By decoupling ground-truth generation from sensor visibility, it provides dense and complete voxel labels ideal for collaborative 3D semantic occupancy prediction.

B. Annotation Statistics

To support tasks with varying ranges and resolutions, we generate dense 3D semantic voxel annotations covering a spatial extent of $100 \times 100 \times 7m^3$ centered on each ego vehicle. The size of each voxel is $0.1 \times 0.1 \times 0.1m^3$, resulting in a total voxel resolution of $1000 \times 1000 \times 70$. All voxels are referenced to vehicle's coordinate, with axes bounded by $x \in [-50, 50]m$ (left to right), $y \in [-50, 50]m$ (front to back), and $z \in [-2, 5]m$ (bottom to top). Additionally, we provide preprocessing tools and the source code for the developed sensor, allowing users to adjust spatial extent or resolution for custom tasks.

To validate the accuracy and completeness of our annotations, we conduct a visual comparison against labels generated from LiDAR scans using the annotation way in SurroundOcc [35]. Figure 2 illustrates representative V2V scenarios, highlighting voxel labels produced by both approaches. Figure 2 illustrates representative V2V scenarios, highlighting voxel labels produced by both approaches. As illustrated, our pipeline yields significantly denser and more complete semantic voxel annotations, especially in occluded or long-range regions where LiDAR-based annotations suffer

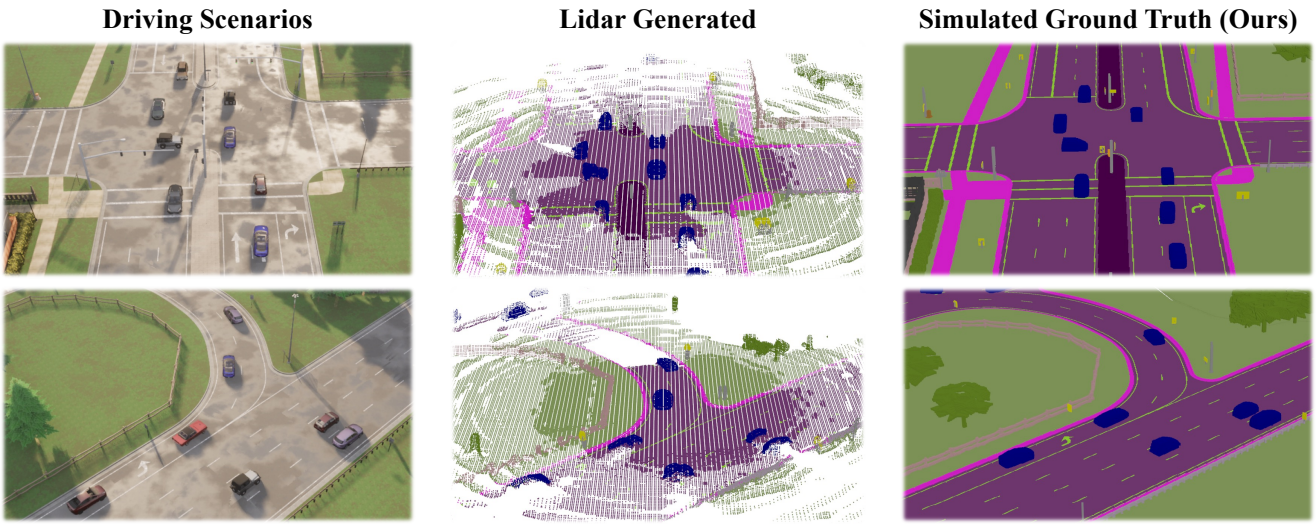

Fig. 2: **Illustration of the V2V scenarios in Carla and the corresponding data collection results.** Left: The screen shot of two V2V scenarios in Carla based on the settings in OPV2V. Mid: LiDAR generated 3D semantic voxel annotations. Right: The annotations collected by our developed 3D semantic voxel sensor.

from sparsity and missing data.

Co3SOP inherits the multi-agent traffic scenarios of OPV2V [13], retaining its training, validation, and testing splits. Distinctively, we treat each vehicle in a scene as an independent training target, allowing every agent to serve as the ego vehicle while dynamically collaborating with its surrounding vehicles. This formulation not only maximizes data utilization, but also enables flexible many-to-many collaboration configurations across agents.

Moreover, to enable fine-grained evaluation of collaborative perception, particularly the ability to recover blind spots beyond the ego vehicle's field of view, we depart from the common practice of marking unobserved voxels as 'unknown'. Instead, we leverage CARLA's semantic engine to assign each voxel a category label, including 'empty' for free space, eliminating the unknown labels. As a result, Co3SOP includes 24 semantic categories (including empty), offering broader class diversity than existing 3D semantic occupancy prediction datasets.

IV. Co3SOP Baseline

To benchmark our proposed Co3SOP dataset, we propose a Collaborative 3D Occupancy Prediction Baseline (Co3SOP-Base), designed to efficiently integrate multi-agent observations and spatial knowledge into unified voxel-based predictions.

A. Overall Structure

The overall structure of Co3SOP-Base is illustrated in fig. 3. The framework first uses a shared backbone network to extract multi-scale image features $\{X_i^l\}_{l=1}^L, X_i^l \in R^{N\times H\times W\times C}$ for each vehicle V_i, where L represent the scales amount of image features and i is vehicle id. These 2D features are then lifted into the 3D voxel space via an image deformable cross-attention module. Once voxel features $F_i \in R^{X\times Y\times Z\times C}$ are obtained for each vehicle, collaborative fusion is enabled by sharing features among connected vehicles (CVs). Upon receiving voxel features from neighboring vehicles, the ego vehicle performs 3D affine transformation to align these features into its own coordinate frame, ensuring proper spatial alignment. Then, the ego vehicle fuses the aligned features with its own voxel representation using a voxel deformable cross-attention mechanism, modulated by a hybrid attention mask that integrates warping and confidence priors. Finally, a prediction head with 3D convolution predicts the voxel-level semantic occupancy map. The details of the key modules are as follows.

Image Backbone. We adopt ResNet101-DCN [36] as the image backbone to extract multi-scale image features from multi-view camera images. The extracted features are further refined using a Feature Pyramid Network (FPN) [37] to aggregate contextual information across scales.

Image To Voxel Transformation. To integrate the 2D image feature into the 3D space, we apply a image deformable cross-attention as in SurroundOcc [35]. This image deformable cross-attention mechanism aggregates multi-view 2D features and learns to map them into the corresponding 3D voxel locations, by considering geometric relationships, ensuring that spatial consistency is maintained when transitioning from 2D to 3D space:

$$
\begin{aligned}
&MSDeformAttn(q,p,x) = \\
&\quad \sum_{m=1}^{N_{head}} W_m[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlk}\cdot W_m^{'} x_i^l(\phi_l(p_q)+\Delta p_{qmlk})] \\
&f_i = MSDeformAttn(q_i, p_q, \{x_i^l\}_{l=1}^L) \qquad (1)
\end{aligned}
$$

where $q_i \in Q_i$ is the corresponding position of query, Q_i is a set of learnable voxel query, W_m and $W_m^{'}$ are learnable weights, A_{mlk} is attention weight, p_q is the reference point of

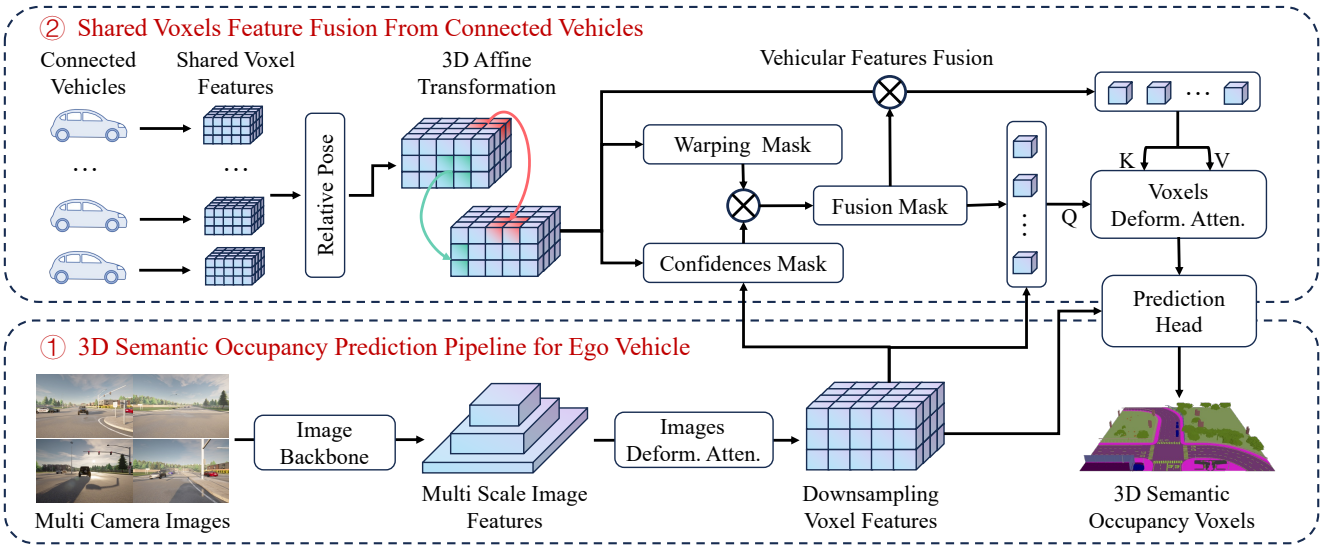

Fig. 3: The Collaborative 3D Semantic Occupancy Prediction Baseline (Co3SOP-Base) consists of two pipelines: (1) ego prediction pipeline, including image backbone, image deformable attention and prediction head; (2) V2V feature fusion pipeline, including 3D affine transformation and mask voxel deformable attention.

the voxel on the image and ϕ_l indicate the rescale sampling function.

Collaborative Vehicle Spatial Alignment. Before multi-agent fusion, voxel features from neighboring agents are transformed into the ego vehicle's coordinate frame using a 3D affine transformation module inspired by Spatial Transformer Networks [38]. The transformation matrix is computed from relative poses between agents, ensuring that voxel features from different coordinates are properly aligned for subsequent attention-based fusion:

$$F_j^{'} = AffineGrid(F_j, T_{ji}^{4\times4}) \tag{2}$$

where $AffineGrid$ is a projective transformation function and $T_{ji}^{4\times4}$ is the pose transformation matrix from vehicle j to vehicle i.

Confidence Masked Feature Fusion. To enable robust and adaptive fusion, each vehicle maintains a learnable confidence mask that estimates the voxel-level reliability of its features. In addition, a binary warping mask is generated during feature alignment to indicate valid projection areas. These two masks are combined to form a hybrid attention mask, which modulates the voxel deformable cross-attention. This design allows the ego vehicle to selectively aggregate features from its collaborators based on both alignment accuracy and confidence, suppressing unreliable or noisy regions. The voxel deformable attention dynamically samples neighboring points across agents, guided by learned offsets and the hybrid attention mask:

$$f_i^{'} = MSDeformAttn(f_i, p_{ij}, \{f_j\}_{j=1}^{J}) \tag{3}$$

where p_{ij} is the reference point that $\{f_j\}_{j=1}^{J}$ has higher confidence than f_i.

Prediction Head. The final stage of Co3SOP-Base is the prediction head, which applies a series of 3D convolutions and 3D deconvolutions to progressively upsample the fused voxel features to the original resolution and output the 3D semantic occupancy results.

B. Loss

To supervise collaborative 3D occupancy prediction, we combine a voxel-level cross-entropy loss for semantic classification, a scene-class affinity loss [18] to encourage consistency between semantic and geometric features, and a confidence loss that regularizes the learned confidence mask for inter-agent feature consistency.

V. Experiments

A. Dataset and Metrics

To evaluate the performance of collaborative 3D semantic occupancy prediction under varying spatial conditions, we define three perception range settings in our benchmark: $25.6 \times 25.6 \times 4.8\,m^3$ range with $0.1\,m$ voxel size, $51.2 \times 51.2 \times 4.8\,m^3$ range with $0.2\,m$ voxel size, and $76.8 \times 76.8 \times 4.8\,m^3$ range with $0.3\,m$ voxel size. These settings allow us to assess how collaborative perception improves scene understanding as the coverage area increases, particularly in occluded or distant regions where a single vehicle's sensing capability is limited. For evaluation, we adopt Intersection-over-Union (mIoU), including IoU for each class individually, as well as the mean IoU (mIoU) across all semantic categories.

B. Benchmark Methods

To establish a fair comparison under our benchmark, we evaluate four representative 3D semantic occupancy prediction models from two modalities. For LiDAR-based methods, we select SSCNet [14] and LMSCNet [15]; for camera-based methods, we include SurroundOcc [35] and OccFormer [20]. These models are chosen for their widespread adoption and

TABLE II: **3D semantic occupancy prediction results** on our Co3SOP dataset. We report the mIoU of some state-of-art methods and our baseline model under three tasks with different ranges, i.e., $25.6 \times 25.6 \times 4.8m^3$, $51.2 \times 51.2 \times 4.8m^3$ and $76.8 \times 76.8 \times 4.8m^3$, as well as the IoUs for some object class. The top performances are highlighted in bold.

Method	SSCNet			LMSCNet			OccFormer			SurroundOcc			**Co3SOP-Base**		
Modality	Lidar			Lidar			Camera			Camera			Camera		
Range	25.6m	51.2m	76.8m	25.6m	51.2m	76.8m	25.6m	51.2m	76.8m	25.6m	51.2m	76.8m	25.6m	51.2m	76.8m
mIoU	13.21	9.58	10.04	24.92	20.35	17.62	29.48	25.41	24.12	28.71	25.76	24.68	**30.04**	**27.50**	**27.00**
Buildings	1.84	0.17	0.19	8.67	3.09	1.79	**11.63**	**11.93**	**13.43**	10.63	7.57	6.91	10.05	8.19	9.15
Fences	0.16	1.48	0.41	**22.27**	**18.01**	9.26	14.17	11.60	11.04	11.06	13.28	8.84	12.37	13.70	**11.38**
Other	0.00	0.00	**16.18**	0.00	0.00	0.00	0.00	0.35	2.74	0.00	**1.77**	12.50	0.00	0.52	6.83
Poles	3.60	0.14	0.00	**29.57**	**24.95**	**17.92**	19.67	12.62	10.17	17.22	13.51	4.78	20.02	16.16	7.12
Roadlines	0.00	0.16	0.00	2.57	0.57	0.00	**39.64**	22.10	15.53	26.78	22.13	14.09	38.43	**29.12**	**16.08**
Roads	0.23	25.88	0.14	86.70	75.84	67.99	87.40	75.30	73.69	86.87	79.53	74.87	**89.24**	**82.35**	**80.28**
Sidewalks	19.22	9.57	20.28	42.24	48.66	53.27	45.32	**51.41**	**59.51**	**46.61**	45.23	55.30	46.12	42.92	55.47
Vegetation	41.43	30.89	22.91	43.77	34.90	23.91	42.78	**39.77**	**35.30**	44.92	35.60	29.08	**46.36**	36.30	34.26
Vehicles	71.73	48.09	39.35	**85.35**	**75.63**	**62.94**	75.70	51.25	33.35	75.95	52.34	29.20	80.55	66.48	50.98
Walls	0.26	0.49	0.18	9.97	10.39	10.08	**13.41**	**15.53**	11.55	12.37	12.92	10.03	12.11	13.99	**12.70**
Trafficsigns	0.00	0.00	0.00	**18.19**	0.02	0.04	9.73	7.68	2.49	17.27	**11.72**	6.23	11.16	9.16	**10.02**
Ground	37.73	0.08	22.18	62.68	31.81	20.59	**67.08**	**57.79**	64.75	53.80	52.90	64.77	55.84	51.96	**68.88**
Bridge	0.00	0.03	0.07	0.00	0.00	0.00	0.00	**2.95**	**5.45**	0.00	2.32	3.74	0.00	2.26	4.39
Guardrail	8.22	12.72	10.21	12.02	6.07	3.37	35.53	41.41	36.90	48.49	42.17	39.65	**53.23**	**48.54**	**44.89**
Trafficlight	0.25	0.00	0.00	0.00	0.00	0.00	**5.43**	**3.75**	0.11	2.11	2.03	1.10	1.27	3.30	**2.42**
Terrain	26.41	2.74	17.01	36.11	36.93	33.43	**86.95**	53.91	53.05	76.58	75.08	**75.08**	82.93	**80.75**	74.43

strong performance in prior benchmarks. We adopt their official implementations and default configurations, modifying only the data loaders and batch sizes to ensure compatibility with the Co3SOP dataset.

C. Implementation Details

For the image backbone, we adapted the weights from FCOS3D [39] as the pretrained weights and set the number of output levels to 4. And these image features from stages 1, 2, and 3 are then fed into the FPN to obtain 4 levels of multi-scale features. For the image and voxel deformable attention, we set the number of layers as 3 and 1. For intermediate voxel features, we apply a downsampling rate of $\frac{1}{4}$ and upsample them in the prediction head. For the collaborative setting, we set the maximum number of collaborating agents to 6 per ego vehicle, based on number of vehicles in OPV2V. During the training process, we apply multi-scale supervision for the multi-scale outputs from prediction head and image augmentation as in SurroundOcc [35]. All experiments are conducted on 8 RTX 4090 GPUs.

D. Benchmark Analysis

The performance of the benchmark methods, SSCNet, LMSCNet, SurroundOcc, and OccFormer, along with our proposed Co3SOP-Base, is presented in table II, which reports mean IoU (mIoU) as well as class-wise IoUs across three perception ranges.

Among the four selected methods, OccFormer achieves the highest mIoU in the $25.6, m$ range (29.48), while SurroundOcc performs best in the $51.2, m$ and $76.8, m$ ranges (25.76 and 24.68, respectively). Notably, both camera-based methods consistently outperform LiDAR-based methods in overall mIoU. However, LiDAR-based models still demonstrate strengths in certain object classes particularly in small-range such as Vehicles and Poles.

Compared with the four selected methods, our proposed Co3SOP-Base consistently achieves the best overall performance across all three perception ranges, with mIoUs of 30.04 ($25.6m$), 27.50 ($51.2m$), and 27.00 ($76.8m$). Compared to the single vehicle methods, our proposed baseline gaines greater improvements as the prediction range increases, especially for object classes such as vehicles, guardrail and roadlines. These results highlight the effectiveness of collaborative voxel fusion, especially as the spatial extent increases.

E. Ablation Study

Effect of Collaborative Feature Fusion. We investigate the impact of inter-agent collaboration by comparing performance with and without V2V feature fusion across all three perception ranges. As shown in table III, collaboration consistently improves mIoU across three ranges. The gains are most significant in the large-range setting, where agents have broader visibility overlap and benefit more from complementary viewpoints. Object-level analysis reveals that collaboration notably enhances performance for large or occlusion-prone classes such as Vehicles (e.g., 34.43 $\rightarrow$ 50.98 IoU at $76.8m$) and Poles (e.g., 10.34 $\rightarrow$ 16.16 IoU at $51.2m$). However, for thin or rare classes like Trafficlight, collaboration yields little to no improvement—and in some cases even slight drops (e.g., Trafficlight: 3.35 $\rightarrow$ 3.30 at $51.2m$).

To further explore the effect of collaboration scale, we evaluate performance under varying numbers of collaborating vehicles. Collaborating vehicles are selected based on

TABLE III: Ablation study on the impact of collaboration in 3D semantic occupancy prediction.

Range	Collaboration	Buildings	Fences	Other	Poles	Roadlines	Roads	Sidewalks	Vegetation	Vehicles	Walls	Trafficsigns	Ground	Bridge	Guardrail	Trafficlight	Terrain	mIoU
$25.6m$	×	9.96	12.56	0.00	18.73	36.19	88.53	44.69	45.51	77.53	11.13	11.10	55.08	0.00	48.61	1.50	82.49	29.36
	✓	10.05	12.37	0.00	20.02	38.43	89.24	46.12	46.36	80.55	12.11	11.16	55.84	0.00	53.23	1.27	82.93	30.04
$51.2m$	×	7.70	12.60	2.54	10.34	26.17	80.03	41.73	32.79	54.78	13.25	8.53	46.76	2.23	42.83	3.35	78.87	25.69
	✓	8.19	13.70	0.52	16.16	29.12	82.35	42.92	36.30	66.48	13.99	9.16	51.96	2.26	48.54	3.30	80.75	27.50
$76.8m$	×	7.32	9.88	12.04	4.29	14.55	75.54	53.53	31.18	34.32	10.54	7.41	62.58	4.02	40.02	2.18	71.54	24.81
	✓	9.15	11.38	6.83	7.12	16.08	80.28	55.47	34.26	50.98	12.70	10.02	58.88	4.39	44.89	2.42	74.43	27.00

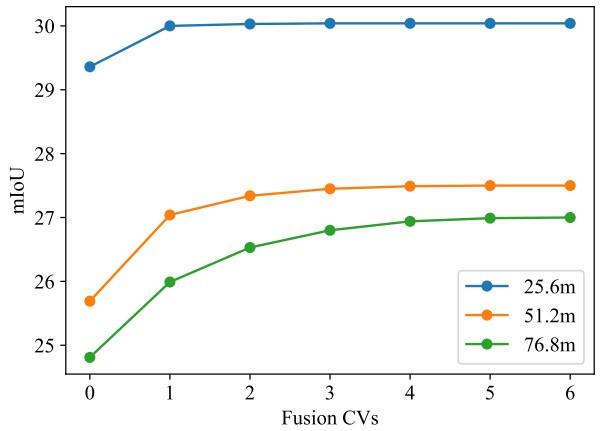

Fig. 4: Ablation study on the number of collaborating vehicles used for feature fusion.

spatial proximity, sorted by distance to the ego vehicle. As shown in fig. 4, performance gains from collaboration mainly come from 1–2 nearby agents. At shorter ranges ($25.6m$), only the closest collaborator yields improvement, while larger-range settings ($51.2m$, $76.8m$) benefit from additional agents due to increased spatial coverage and complementary viewpoints.

Robustness on Pose Noise We further assess the robustness of Co3SOP-Base under localization uncertainty by injecting Gaussian pose noise into the relative transformations used for inter-agent feature alignment. As shown in table IV, increasing the mean offset from $\mu = 0.1m$ to $0.5m$ (with fixed standard deviation $\sigma = 0.02$) leads to a gradual mIoU degradation across all ranges. This decline illustrates that misalignment in inter-agent transformations can affect collaborative feature quality. However, the performance remains relatively stable under moderate noise, suggesting that our warping-based spatial alignment and confidence-aware fusion retain reasonable robustness in the presence of pose perturbations.

VI. Conclusion and Future Work

We presented Co3SOP, a synthetic benchmark for collaborative 3D semantic occupancy prediction, built upon high-fidelity simulation and comprehensive voxel annotations.

TABLE IV: Ablation study on the impact of pose noise on collaborative 3D semantic occupancy prediction.

Noise	Range		
	25.6m	51.2m	76.8m
$\mu = 0.0, \sigma = 0.00$	30.04	27.50	27.00
$\mu = 0.1, \sigma = 0.02$	29.95	27.38	26.96
$\mu = 0.2, \sigma = 0.02$	29.78	27.13	26.83
$\mu = 0.3, \sigma = 0.02$	29.59	26.82	26.64
$\mu = 0.4, \sigma = 0.02$	29.41	26.50	26.39
$\mu = 0.5, \sigma = 0.02$	29.30	26.20	26.15

Unlike prior datasets which are limited by LiDAR sparsity, Co3SOP leverages a custom-designed voxel annotation pipeline to provide complete and fine-grained semantic voxel labels, enabling robust evaluation of multi-agent perception systems. We further proposed Co3SOP-Base, a baseline framework incorporating confidence and alignment aware masked voxel deformable attention for multi-agent feature fusion. Extensive experiments demonstrate the effectiveness of V2V collaboration in improving perception performance. This work addresses the critical gap in collaborative 3D voxel-level understanding by providing both a benchmark and a baseline, paving the way for safer and more robust autonomous driving systems.

The primary limitation of this work lies in the the reliance on purely synthetic data. Future work could integrate temporally asynchronous modeling, communication cost considerations and sim-to-real transfer evaluation.

Acknowledgment

This work was supported by the JST ASPIRE Program (Grant Number JPMJAP2325), JST CRONOS (Grant Number JPMJCS24K8), JST SPRING (Grant Number JPMJSP2108) and JSPS Grant-in-Aid for Early-Career Scientists (Grant Number JP25K21195).

References

[1] S. Liu, C. Gao, Y. Chen, X. Peng, X. Kong, K. Wang, R. Xu, W. Jiang, H. Xiang, J. Ma, *et al.*, "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," *arXiv preprint arXiv:2308.16714*, 2023.

[2] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "V2x cooperative perception for autonomous driving: Recent advances and challenges," *arXiv preprint arXiv:2310.03525*, 2023.

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

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

[5] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] H. Xu, J. Chen, S. Meng, Y. Wang, and L.-P. Chau, "A survey on occupancy perception for autonomous driving: The information fusion perspective," *arXiv preprint arXiv:2405.05173*, 2024.

[7] Y. Shi, K. Jiang, J. Li, J. Wen, Z. Qian, M. Yang, K. Wang, and D. Yang, "Grid-centric traffic scenario perception for autonomous driving: A comprehensive review," *arXiv preprint arXiv:2303.01212*, 2023.

[8] Y. Zhang, J. Zhang, Z. Wang, J. Xu, and D. Huang, "Vision-based 3d occupancy prediction in autonomous driving: a review and outlook," *arXiv preprint arXiv:2405.02595*, 2024.

[9] S. Teufel, J. Gamerdinger, J.-P. Kirchner, G. Volk, and O. Bringmann, "Collective perception datasets for autonomous driving: A comprehensive review," *arXiv preprint arXiv:2405.16973*, 2024.

[10] M. Yazgan, M. V. Akkanapragada, and J. M. Zöllner, "Collaborative perception datasets in autonomous driving: A survey," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2269–2276.

[11] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

[12] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2583–2589.

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

[15] L. Roldão, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 111–119.

[16] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3cnet: A sparse semantic scene completion network for lidar point clouds," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 2148–2161.

[17] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109.

[18] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3991–4001.

[19] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9087–9098.

[20] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 9433–9443.

[21] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss, "Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset," *The International Journal on Robotics Research*, vol. 40, no. 8-9, pp. 959–967, 2021.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[23] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[24] Y. Li, S. Li, X. Liu, M. Gong, K. Li, N. Chen, Z. Wang, Z. Li, T. Jiang, F. Yu, *et al.*, "Sscbench: Monocular 3d semantic scene completion benchmark in street views," *arXiv preprint arXiv:2306.09001*, 2023.

[25] Y. Zhang, J. Li, K. Luo, Y. Yang, J. Han, N. Liu, D. Qin, P. Han, and C. Xu, "V2vssc: A 3d semantic scene completion benchmark for perception with vehicle to vehicle communication," *arXiv preprint arXiv:2402.04671*, 2024.

[26] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.

[27] J. Mei, A. Z. Zhu, X. Yan, H. Yan, S. Qiao, L.-C. Chen, and H. Kretzschmar, "Waymo open dataset: Panoramic video panoptic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 53–72.

[28] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[29] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 668–22 677.

[30] R. Hao, S. Fan, Y. Dai, Z. Zhang, C. Li, Y. Wang, H. Yu, W. Yang, J. Yuan, and Z. Nie, "Rcooper: A real-world large-scale dataset for roadside cooperative perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 347–22 357.

[31] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[32] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, *et al.*, "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*, 2023, pp. 13 712–13 722.

[33] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[34] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun, *et al.*, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5486–5495.

[35] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 21 729–21 740.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[38] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[39] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.