

MCOP: Multi-UAV Collaborative Occupancy Prediction

Zefu Lin^{1,2,3} Wenbo Chen^{2,3} Xiaojuan Jin^{2,3} Yuran Yang^{5,6} Lue Fan^{2,3}

Yixin Zhang⁶ Yufeng Zhang¹ Zhaoxiang Zhang^{1,2,3,4 *}

¹ University of Chinese Academy of Sciences (UCAS)

² Institute of Automation, Chinese Academy of Sciences (CASIA)

³ New Laboratory of Pattern Recognition (NLPR)

⁴ State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS)

⁵ Beijing University of Posts and Telecommunications (BUPT) ⁶ Tencent

{linzefu2022, wenbo.chen, xiaojuan.jin, lue.fan, zhaoxiang.zhang}@ia.ac.cn

Abstract

Unmanned Aerial Vehicle (UAV) swarm systems necessitate efficient collaborative perception mechanisms for diverse operational scenarios. Current Bird's Eye View (BEV)-based approaches exhibit two main limitations: bounding-box representations fail to capture complete semantic and geometric information of the scene, and their performance significantly degrades when encountering undefined or occluded objects. To address these limitations, we propose a novel multi-UAV collaborative occupancy prediction framework. Our framework effectively preserves 3D spatial structures and semantics through integrating a Spatial-Aware Feature Encoder and Cross-Agent Feature Integration. To enhance efficiency, we further introduce Altitude-Aware Feature Reduction to compactly represent scene information, along with a Dual-Mask Perceptual Guidance mechanism to adaptively select features and reduce communication overhead. Due to the absence of suitable benchmark datasets, we extend three datasets for evaluation: two virtual datasets (Air-to-Pred-Occ and UAV3D-Occ) and one real-world dataset (GauUScene-Occ). Experiments results demonstrate that our method achieves state-of-the-art accuracy, significantly outperforming existing collaborative methods while reducing communication overhead to only a fraction of previous approaches.

1. Introduction

Unmanned Aerial Vehicles (UAVs) are increasingly used in applications such as smart cities [1], traffic management, and emergency response [9]. These applications require advanced environmental perception capabilities that single-UAV systems inherently lack due to their limited field of

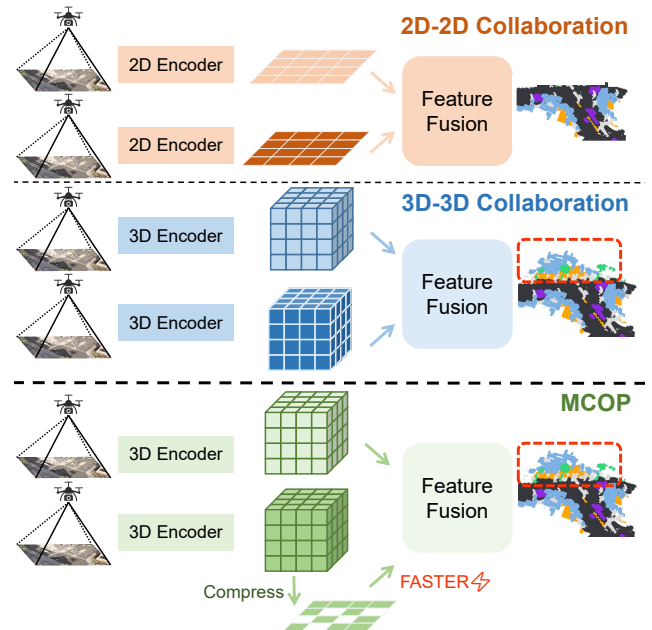


Figure 1. **Comparison of Multi-UAV collaborative occupancy prediction and other collaborative methods.** 2D-to-2D method lacks height information and cannot effectively reconstruct the 3D details of the scene. 3D-to-3D method requires transmitting high-dimensional occupancy features, which demands high bandwidth and affects real-time performance. MCOP compresses occupancy features, balancing transmission rate and prediction quality.

view and susceptibility to occlusions. To address these challenges, multi-UAV collaborative perception has emerged as a promising solution by integrating observations from multiple viewpoints to enhance scene understanding.

Current multi-UAV perception systems typically project image features into a unified Bird's-Eye-View (BEV) coordinate system for 3D object detection [6, 32, 38, 44]. While effective for identifying specific objects, these approaches

*Corresponding author

fail to capture the rich geometric and semantic details of the environment, such as irregularly shaped obstacles or partially occluded structures. Furthermore, BEV-based features lack altitude information, which is particularly crucial for UAVs due to their reliance on 3D spatial awareness. To overcome these limitations, recent research has explored the use of 3D occupancy prediction [30], which represents the environment as a voxel grid encoding both occupancy status and semantic categories. Unlike bounding box-based methods, occupancy prediction provides a comprehensive understanding of the 3D scene, including free space, occupied regions, and undefined obstacles. However, extending occupancy prediction to multi-UAV collaborative scenarios introduces a new set of challenges. UAVs typically operate at altitudes $10\times$ higher than ground-based autonomous vehicles, requiring them to perceive a broader range of scenes—from ground surfaces to buildings and aerial objects. This significantly expands the feature space for occupancy representation, making real-time processing and communication infeasible with existing methods [29].

In this paper, we propose Multi-UAV Collaborative Occupancy Prediction (MCOP), a vision-centric framework designed to address the unique challenges of UAV-based 3D scene understanding. Our approach leverages the rich geometric and semantic information provided by occupancy prediction while overcoming the computational and communication bottlenecks associated with multi-UAV collaboration. The core of MCOP lies in its novel visual feature representation and fusion mechanisms, which are specifically designed for UAV viewpoints. First, we introduce the Spatial-Aware Feature Encoder, which transforms RGB images into 3D occupancy features using a combination of Voxel-Image Attention and Cross-Voxel Attention. This encoder effectively captures detailed scene geometry and semantics without relying on depth sensors, making it suitable for resource-constrained UAV platforms. To address the high-dimensionality of occupancy features, we propose Altitude-Aware Reduction, a compression mechanism that retains critical height information while reducing feature dimensions. This is achieved by encoding vertical pillars into 2D BEV representations, significantly reducing communication overhead without sacrificing perceptual accuracy. Furthermore, we develop Dual-Mask Perceptual Guidance, a dynamic feature selection mechanism that identifies and transmits only the most relevant visual information across UAVs. By leveraging support masks (high-confidence regions) and request masks (low-confidence regions), this module minimizes redundant data transmission while ensuring robust perception in occluded or complex scenes. Finally, the Cross-Agent Feature Integration module fuses local and received features into a unified 3D occupancy representation, enabling comprehensive scene understanding across multiple UAVs.

Because 3D occupancy labeling is expensive, no public dataset currently supports multi-UAV collaborative semantic occupancy prediction. To address this gap, we extend three datasets for our evaluation: two CARLA-based virtual datasets, Air-to-Pred-Occ [35] and UAV3D-Occ [42], and one real-world dataset, GauUScene-Occ [39]. We enrich each with 3D occupancy annotations, thereby filling a crucial gap in UAV collaborative perception research. Inspired by [30], we employ a streamlined method to derive suitable occupancy ground truth for these aerial scenarios.

Experimental results demonstrate that collaborative perception consistently outperforms single-UAV perception in semantic occupancy prediction, benefiting from enhanced spatial coverage and information sharing. Comparative analysis with adapted autonomous driving approaches BEVDet [14] and PanoOcc [34] shows that our method achieves higher mIoU on all evaluated datasets with significantly reduced communication overhead (0.23 MB vs. 17.50 MB and 19.14 MB, respectively).

Contributions Our key contributions are:

- **Occupancy-Based Multi-UAV Perception Framework.** To our knowledge, we propose the first collaborative occupancy prediction framework for multi-UAV systems. Our method addresses key limitations of BEV-based approaches by effectively preserving rich semantic and geometric information including occluded objects.
- **High-Efficiency Collaboration Strategy.** Altitude-Aware Reduction and Dual-Mask Perceptual Guidance significantly lower communication overhead while preserving essential 3D features, thus supporting real-time collaboration among UAVs.
- **Enriched Collaborative Datasets.** We extend three datasets with occupancy annotations, offering a new benchmark for multi-UAV semantic occupancy prediction and fostering further exploration in aerial 3D perception research.

2. Related Work

2.1. Collaborative Prediction

In multi-agent systems, sharing information across perception nodes (vehicles, infrastructure, etc.) effectively expands a node's field of view and mitigates occlusion-induced degradation [1, 26]. In large-scale scenarios, collaborative perception significantly improves detection accuracy and robustness over individual perception [10].

Collaboration strategies are typically categorized as early, intermediate, or late, based on the fusion stage of sensing modalities [35]. Early collaboration fuses raw data at the input layer [2], maximizing shared content but requiring high bandwidth. Late collaboration merges target predictions at the output [6], conserving bandwidth but often amplifying accumulated noise. Intermediate collaboration,

focusing on feature-level fusion [31], achieves a balanced trade-off between communication cost and accuracy [37].

Information-sharing strategies differ among nodes. Some methods share all data to maximize coverage, at the cost of bandwidth. To reduce redundancy, dynamic communication strategies like Who2com [22] and When2com [21] use attention or scheduling to determine optimal communication timing and partners. Where2comm [12] further selects informative local features based on regional uncertainty.

Feature fusion began with simple operations [45]. F-Cooper [5] uses element-wise max for voxel-level fusion; CoHFF [29] incorporates similarity-based weighting to exploit complementary, low-confidence features. V2VNet [33] applies a variational graph network, while DiscoNet [16] introduces matrix-valued weights for fine-grained attention. Recent transformer-based models such as V2X-ViT [41] and CoBEVT [40] use multi-agent attention for multi-camera fusion. CoCa3D [13] enhances depth prediction using uncertainty to improve cross-view fusion. However, these methods mainly target 2D feature fusion; moving to 3D requires additional mechanisms to preserve real-time performance.

2.2. Occupancy Prediction

Unlike detection-based methods, occupancy prediction estimates the semantic state of each voxel. Reconstructing 3D scenes from visual input demands complete geometry and semantic reasoning, posing challenges due to high dimensionality and data sparsity [3, 25, 36, 43, 45].

To mitigate 2D-to-3D projection uncertainty, FB-BEV [19] uses both forward and backward projections and applies depth-consistency weighting. Addressing height-information loss in standard BEV projection, the TPV family [15, 28] exploits three complementary viewpoints (top, front, side). Alternatively, some methods directly process 3D features. MiLO [24] uses 3D ResNet [11] and FPN [20], PanoOcc [34] merges spatiotemporal voxel queries for detailed 3D information. Voxformer [17] employs sparse voxel queries to index 2D features via camera projection. COTR [23] leverages geometry priors and explicit-implicit transforms to reduce voxel sparsity.

The primary challenge lies in effectively learning high-dimensional and sparse 3D features. Methods based on BEV, TPV, or direct 3D operations each address the core issue of representation sparsity and depth inference from a different angle. This becomes especially difficult when aligning multi-view 2D inputs with 3D space in large-scale or dynamic settings. In multi-UAV cooperative perception, for example, frequent viewpoint shifts, larger feature dimensions, and greater motion variability make stable feature extraction and alignment even more demanding. Nonetheless, coupling occupancy prediction with collabora-

tive strategies—such as leveraging uncertainty or visibility masks to reduce redundant transmissions—can still deliver a refined, complete representation of the environment.

3. Methodology

Our MCOP framework consists of four key modules, namely Spatial-Aware Feature Encoder, Altitude-Aware Reduction, Dual-Mask Perceptual Guidance and Cross-Agent Feature Integration. It achieves efficient inter-UAV collaborative prediction with minimal accuracy cost by transmitting encoded spatial-aware occupancy features.

3.1. Problem Setup

In the multi-UAV collaborative 3D occupancy prediction task, we define the UAV network by a global communication network, represented as an undirected graph $\mathcal{G} = (\mathcal{X}, \mathcal{L})$, where $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$ denotes all UAVs, and $\mathcal{L} = \{L_{ij} \mid i, j \in [1, n]\}$ where L_{ij} denotes the communication links between UAV X_i and X_j . For each UAV X_i , the set of connected UAVs is represented as $\mathcal{N}_i = \{X_j \mid L_{ij} \in \mathcal{L}, j \in [1, n]\}$, where \mathcal{N}_i denotes all the UAVs directly communicating with UAV X_i . UAV X_i takes RGB images $\mathcal{I}_i \in \mathbb{R}^{H \times W \times 3}$ as input, and outputs 3D occupancy prediction $\mathcal{O} \in \mathbb{R}^{X \times Y \times Z}$ with certain semantic categories, and X, Y, Z are dimensions of 3D occupancy voxel space. Inspired by [29], the optimization problem is defined as follows

$$\begin{aligned} \max_{\theta, F} \quad & \sum_{X_i \in \mathcal{X}} g(\Phi_{\theta}(\mathcal{I}_i, \{F_{j \rightarrow i} \mid X_j \in \mathcal{N}_i\}), \mathcal{O}_i^{gt}), \\ \text{s.t.} \quad & \sum_{X_i \in \mathcal{X}} \sum_{X_j \in \mathcal{N}_i} |F_{j \rightarrow i}| \leq B, \quad (1) \end{aligned}$$

where Φ_{θ} represents the model parameterized by θ , $F_{j \rightarrow i}$ denotes the features transmitted from UAV X_j to UAV X_i , and $g(\cdot, \cdot)$ represents the function to evaluate the predicted occupancy against the ground truth \mathcal{O}_i^{gt} . B denotes the dynamic communication volume constraint, which may vary depending on hardware conditions. The optimization objective is to maximize the overall perception effectiveness within the communication upper bound $B \in \mathbb{R}^+$.

3.2. Overall Architecture

This section introduces the overall architecture of MCOP. As illustrated in Fig. 2, each UAV independently takes RGB images as input and uses Spatial-Aware Feature Encoder 3.3, which includes an image backbone and an Occupancy Encoder, to generate 3D occupancy features. To minimize the data required for collaborative communication, these 3D occupancy features are compressed into 2D BEV features via the Altitude-Aware Reduction 3.4. It encodes features based on effective spatial information, thus

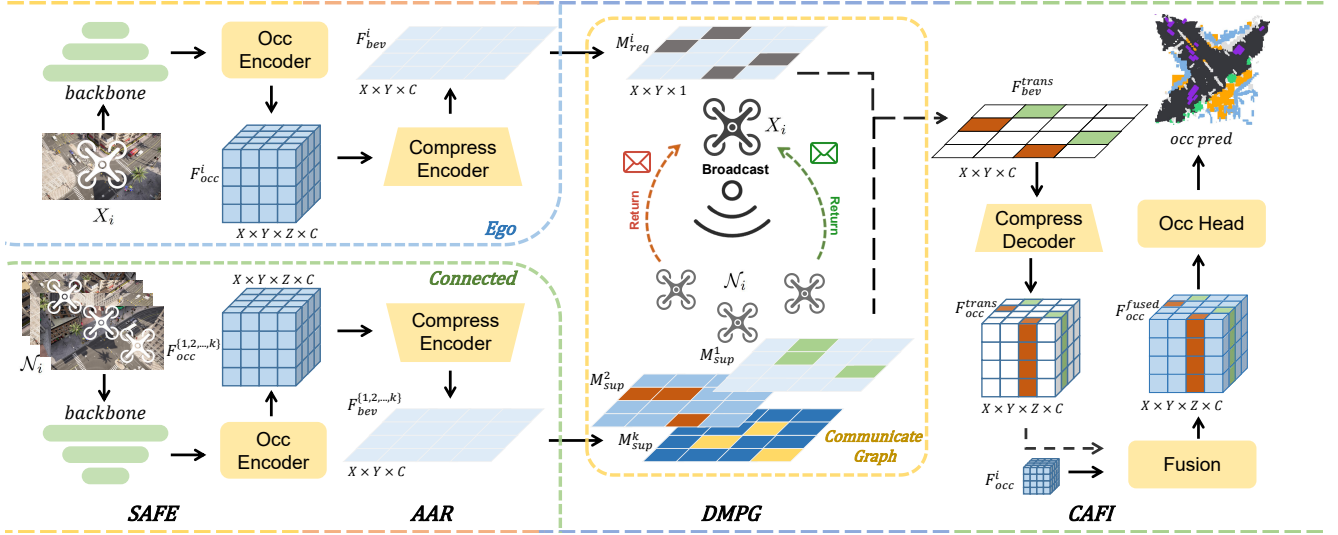


Figure 2. **The overall framework of MCOP.** Each UAV uses an image backbone to extract multi-scale features, which are processed by the *Spatial-Aware Feature Encoder (SAFE)* to generate 3D occupancy features. The *Altitude-Aware Reduction (AAR)* compresses these 3D features into compact 2D BEV representation for efficient communication. *Dual-Mask Perceptual Guidance (DMPG)* coordinates the sharing of relevant information among UAVs based on perception quality. The *Cross-Agent Feature Integration (CAFI)* module fuses local and received features into unified 3D representation. Finally, the Occ Head predicts 3D occupancy segmentation, resulting in comprehensive environmental perception among UAVs.

reducing bandwidth requirements. Up to this stage, each UAV operates independently without interaction.

In the Dual-Mask Perceptual Guidance module 3.5, each UAV assesses the perceptual quality of its local regions based on its 2D BEV features. It then generates two types of masks: a support mask, representing regions with high perceptual confidence, and a request mask, indicating areas with low perceptual confidence that require assistance from other UAVs. During each communication phase, the ego UAV broadcasts its request mask to solicit assistance from other connected UAVs. Connected UAVs project their support masks into the ego UAV’s perception space and compute the intersection with the ego’s request mask to determine the regions requiring collaboration. The connected UAVs then transmit the corresponding compressed feature data for these regions to the ego UAV, enabling collaborative perception. This interaction ensures that the ego UAV receives only the necessary, high-confidence information.

After receiving information from other UAVs, the ego UAV applies Cross-Agent Feature Integration 3.6 to combine its 3D occupancy features with the received 2D features, yielding 3D fused occupancy representation. This fused representation is subsequently used by the task processing head for 3D occupancy segmentation. These modules together enable our method to achieve efficient collaborative perception among multiple UAVs. The following sections describe each module in detail. In the following sections, we provide detailed descriptions of each module.

3.3. Spatial-Aware Feature Encoder

For input RGB images, we first use pretrained backbone network (e.g. ResNet [11]) to extract image features. To capture detailed scene information without relying on depth, we follow [34] and define a set of 3D voxel queries $\mathbf{Q} \in \mathbb{R}^{X \times Y \times Z \times C}$, where C represents the feature channels, and X, Y, Z are the voxel grid dimensions. We propose Voxel-Image Attention to bridge feature extraction and voxel representation, which uses deformable attention [47] to associate each voxel query \mathbf{q} at (x, y, z) with relevant image features. This Voxel-Image Attention (VIA) can be defined as

$$\text{VIA}(\mathbf{q}, f(\mathcal{I}_i)) = \sum_{\eta=1}^{\mathcal{P}_s} \text{DA}(\mathbf{q}, \delta(\text{Ref}_{(x,y,z)}^\eta), f(\mathcal{I}_i)), \quad (2)$$

where f is image backbone, \mathcal{P}_s is the number of sampling points per voxel query, and $\delta(\text{Ref}_{(x,y,z)}^\eta)$ denotes the η -th sampling point projected onto the voxel grid using projection matrix δ . DA denotes deformable attention. During this process, the querying paradigm [18] efficiently transforms perspective view features into voxel space representations, reducing computational complexity. We next apply Cross-Voxel Attention (CVA) to establish connections between voxel queries, which is defined as

$$\text{CVA}(\mathbf{q}, \mathbf{Q}) = \sum_{\eta=1}^{\mathcal{P}_r} \text{DA}(\mathbf{q}, \text{Ref}_{(x,y,z)}^\eta, \mathbf{Q}), \quad (3)$$

where \mathcal{P}_r is the number of reference points per voxel query.

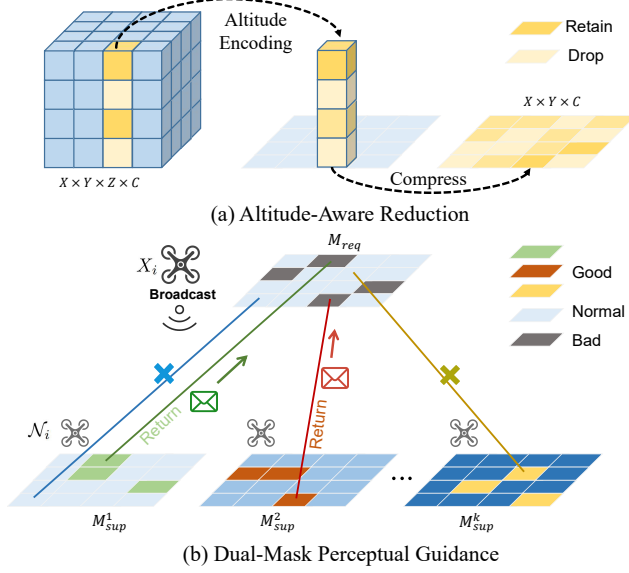


Figure 3. **Illustration of Altitude-Aware Reduction and Dual-Mask Perceptual Guidance.** (a) demonstrates how a pillar in the 3D occupancy feature is compressed into a grid in the 2D BEV feature. (b) illustrates the details of DMPG, where the ego UAV first broadcasts its request mask within the network, and then the connected UAVs select and transmit high-quality features that are needed by the ego UAV.

These operations allow voxel queries to interact both with image pixels and with each other, enriching the geometric and semantic content. Finally, we obtain the occupancy feature $F_{occ}^i \in \mathbb{R}^{X \times Y \times Z \times C}$.

3.4. Altitude-Aware Reduction

In multi-UAV perception system, common collaboration strategy is to transmit encoded image features to share environmental information. This approach is more efficient than transmitting raw images and outperforms sharing post-processed perception results. However, for occupancy prediction, the encoder output is 3D features, and transmitting 3D features directly still requires significant bandwidth. Given that only about 5% of the spatial areas are occupied, we designed Altitude-Aware Reduction(AAR) to compress 3D occupancy features into 2D BEV features, thereby reducing communication costs.

As shown in Fig. 3(a), we first normalize the 3D features using sigmoid function and apply a threshold θ to filter valid spatial points. For each pillar, we create an altitude embedding along the Z-axis as $[1, 2, 3, \dots, Z - 1]$ and a binary index, where 1 represents a valid point and 0 represents an invalid point to be filtered. Next, we use the index to compute weighted sum of the altitude embedding, and then calculate average altitude value by dividing the number of valid points in the pillar. This reduces the dimensionality of the altitude information. Subsequently, we normalize the

average altitude of each pillar to the range of $[0, 1]$, generating 2D altitude encoding $\mathcal{A}^i \in \mathbb{R}^{X \times Y}$, which represents the altitude information for each 2D grid.

To retain further 3D context, we compute weighted average of the 3D features along the Z-axis and concatenate it with \mathcal{A}^i . The resulting representation is then passed through a 2D convolution layer for additional compression, producing the final compressed 2D feature F_{bev}^i . Comprehensive compression procedure can be formalized as

$$F_{bev}^i = \Psi \left(\frac{1}{|Z|} \sum_z (\mathbf{M}_z^i \odot F_{occ}^i) + \mathcal{A}^i(x, y) \right), \quad (4)$$

where $\sum_z(\cdot)$ denotes the summation over the z axis to reduce the 3D feature into 2D representation. \mathbf{M}_z^i represents a binary mask that identifies valid points in the feature map based on logistic function and threshold. \odot denotes element-wise multiplication between the binary mask \mathbf{M}_z^i and the 3D feature F_{occ}^i . $\mathcal{A}^i(x, y)$ is the altitude encoding for each spatial position (x, y) , retaining height information. $\Psi(\cdot)$ represents the 2D convolution layer for further feature compression after concatenation. F_{bev}^i maintains key spatial information while significantly reducing dimensionality, making it more bandwidth-efficient.

3.5. Dual-Mask Perceptual Guidance

With the integration of AAR, we obtain altitude-aware planar features. In contrast to autonomous driving, where on-board cameras have minimal overlap, UAV mission scenarios involve significant overlap in observation areas between UAVs. Additionally, UAV observations vary in quality due to occlusions or edge distortions, leading to regions of different observational quality. To improve data efficiency, we propose Dual-Mask Perceptual Guidance, shown in Fig. 3(b), with a generation processes for a request mask \mathbf{M}_{req}^i and a support mask \mathbf{M}_{sup}^i . \mathbf{M}_{req}^i identifies areas of poor observation, while \mathbf{M}_{sup}^i selects high-quality regions for data transmission. Each grid in the BEV feature map F_{bev}^i is assigned a quality score based on both distance and feature gradient. Generation of \mathbf{M}_{sup}^i can be formalized as

$$\mathbf{M}_{sup}^i(x, y) = \begin{cases} 1 & \text{if } \alpha \cdot \frac{h}{\sqrt{h^2 + d^2}} + \beta \cdot \frac{|G(x, y)|}{\epsilon} > \xi, \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where α and β are weighting coefficients, h represents the UAV's altitude, d is the horizontal distance to the grid (x, y) , $|G(x, y)|$ is the gradient magnitude, ϵ limits gradient complexity, and ξ represents the quality score threshold. High thresholds restrict data but risk insufficient information; low thresholds transmit more data but can dilute useful features with noise. The impact of quality score threshold is discussed in the ablation study. The underperforming regions in \mathbf{M}_{req}^i are defined as the inverse of \mathbf{M}_{sup}^i .

In each communication round, ego UAV broadcasts request mask M_{req}^i to request high-quality data from neighboring UAVs. Upon receiving M_{req}^i , collaborative UAVs project their support masks $M_{\text{sup}}^{\{1,2,\dots,k\}}$ BEV features and $F_{\text{bev}}^{\{1,2,\dots,k\}}$ into the ego UAV's BEV space. They then apply both M_{req}^i and the projected $M_{\text{sup}}^{\{1,2,\dots,k\}}$ to extract the features for transmission $F_{\text{bev}}^{\text{trans}}$, denoted as

$$F_{\text{bev}}^{\text{trans}} = \left(M_{\text{req}}^i \cap \tau M_{\text{sup}}^{\{1,2,\dots,k\}} \right) \odot \tau F_{\text{bev}}^{\{1,2,\dots,k\}}, \quad (6)$$

where τ denotes the transformation to the ego UAV's reference frame. This ensures that ego UAV receives only essential features, minimizing redundant data transmission.

3.6. Cross-Agent Feature Integration

We propose Cross-Agent Feature Integration (CAFI) to integrate 2D transmitted feature $F_{\text{bev}}^{\text{trans}}$ from connected UAVs with 3D $F_{\text{occ}}^{\{1,2,\dots,k\}}$ of ego UAV. CAFI restores geometric and semantic information through upsampling and fusion, then output semantic occupancy via a task-specific head.

Upsampling. We upsample $F_{\text{bev}}^{\text{trans}}$ to improve feature resolution, followed by a 3D convolution to extend it into the 3D space, generating volumetric features. The resulting feature implicitly retains altitude information, allowing the feature to effectively capture detailed semantic variations in the 3D environment, particularly along the altitude dimension.

Feature Fusion. The upsampled $F_{\text{bev}}^{\text{trans}}$ is then concatenated with ego UAV's 3D feature F_{occ} along the channel dimension. The concatenated features are then processed by a residual 3D convolutional module, yielding fused occupancy feature $F_{\text{occ}}^{\text{fused}}$. To retain sufficient spatial granularity, we use 3D deconvolutions to refine the fused resolution, ensuring high-quality feature representation.

Task Output. For fine-grained semantic scene prediction, we utilize a Multilayer Perceptron (MLP) as the task head for semantic segmentation of the fused 3D features, ultimately producing the final collaborative prediction. Our training approach involves two loss functions. One is semantic segmentation loss L_{seg} , which leverages focal loss to mitigate class imbalance. The second is communication constraint loss L_{com} , which incorporates $L1$ regularization to minimize data transmission overhead. The final optimization objective function is given by: $L = L_{\text{seg}} + \lambda L_{\text{com}}$.

4. Experiment

4.1. Datasets

Due to the lack of a suitable dataset for collaborative UAV occupancy prediction, we incorporate semantic occupancy annotations into three datasets: Air-Co-Pred [35], UAV3D [42], and GauUScene [39]. **Air-Co-Pred** [35] is a Carla-based [8] virtual dataset feature four UAVs monitoring a 100m×100m intersection at a 50m altitude. It

has 32,000 synchronized images (1600×900) split into 170 training and 30 validation scenes. **UAV3D** [42] is a synthetic dataset created with Carla [8] and AirSim [27], covering both urban and suburban environments. Five UAVs fly at 60m altitude, producing 700 training, 150 validation, and 150 test scenes (800×450). We use one town from each virtual dataset for experiments. **GauUScene** [39] is a real-world dataset designed for 3D reconstruction, featuring multiple 1km²-scale scenes with UAV-captured RGB images (5472×3648 resolution), corresponding poses, and point clouds. We use one subset covering 0.908km² ("Russian Building" scene) with UAV flights up to 150m altitude. Since GauUScene [39] is not intended for collaborative perception, we treat its four UAV trajectories in this subset as a single four-UAV cluster to enable cooperative sensing.

Occupancy Annotation. The original Air-Co-Pred [35] and UAV3D [42] datasets only provide 2D and 3D bounding box annotations for vehicles. To facilitate occupancy prediction, we generate additional occupancy annotations for these datasets. Specifically, for Air-Co-Pred-Occ and UAV3D-Occ, we first export mesh maps from CARLA [8], and then annotate semantic labels using a 3D point cloud annotation tool. For GauUScene-Occ [39], we follow the methodology in Occ3D [30] by reconstructing meshes and assigning corresponding semantic labels. All annotations are further refined via ray-casting to realistically simulate occlusions from a single UAV's viewpoint (e.g., objects obscured by walls remain invisible). Final occupancy annotations include seven semantic categories: *free*, *others*, *ground*, *building*, *vegetation*, *vehicle*, and *urban road*. Here, *free* represents unoccupied space as the complement of occupied regions, while *others* denotes objects without specific semantic labels.

Evaluation metrics. Following the evaluation approach for semantic occupancy prediction in autonomous driving, we use Intersection over Union (IoU) as the evaluation metric. This involves computing IoU for each class and the mean IoU (mIoU) across all classes.

4.2. Experiment Settings

Implementation Details. For voxelization, Air-Co-Pred-Occ [35] and UAV3D-Occ [42] use a 0.4m³ voxel size, while GauUScene-Occ [39] adopts a coarser voxel size of 2m³ due to the larger observation space. Air-to-Pred-Occ [35] and GauUScene-Occ [39] each involve four UAVs, while UAV3D-Occ [42] uses five. Each UAV covers an observation range with overlapping regions for enhanced perception. We employ a ResNet101-DCN [7] backbone and FPN [20] at four scales (1/8, 1/16, 1/32, 1/64). Dual-Mask Perceptual Guidance compresses features via two 2D convolutions with a 0.8 quality threshold. Cross-Agent Feature Integration uses hierarchical 3D convolutions for upsampling, and our segmentation head applies two MLP layers

Dataset	Type	Method	Image Size	Co.	Range (m ²)	Height (m)	CV(MB) ↓	mIoU ↑
Air-to-Pred-Occ	Simulated	BEVDet†	1600×900	×	100×100	50	-	7.46
		PanoOcc	1600×900	×	100×100	50	-	40.82
		BEVDet‡	1600×900	✓	100×100	50	17.50	12.29
		PanoOcc‡	1600×900	✓	100×100	50	19.14	41.96
		MCOP (Ours)	1600×900	✓	100×100	50	0.23	46.41
UAV3D-Occ	Simulated	BEVDet†	800×450	×	112×112	60	-	8.21
		PanoOcc	800×450	×	112×112	60	-	43.48
		BEVDet‡	800×450	✓	112×112	60	17.50	12.09
		PanoOcc‡	800×450	✓	112×112	60	19.14	44.73
		MCOP (Ours)	800×450	✓	112×112	60	0.23	47.89
GauUScene-Occ	Real	BEVDet†	5472×3648	×	500×500	150	-	7.27
		PanoOcc	5472×3648	×	500×500	150	-	40.43
		BEVDet‡	5472×3648	✓	500×500	150	17.50	11.18
		PanoOcc‡	5472×3648	✓	500×500	150	19.14	42.69
		MCOP (Ours)	5472×3648	✓	500×500	150	0.23	42.92

Table 1. **Experimental results of different methods on various datasets.** Co. represents whether collaborative perception is applied. Range represents the observation area of UAV, Height refers to the UAV’s flight altitude, and CV denotes the communication volume, which is the data transmission cost per communication instance, measured in MB. † indicates that BEV features are converted into occupancy features using the FlashOcc [43] method for a fair comparison. ‡ denotes the addition of a collaboration module following the Where2comm [12] method. Our method achieves the highest mIoU and the lowest communication volume.

(hidden size 128) with softplus [46] activation.

Training. Training is conducted on eight NVIDIA A6000 GPUs, with a batch size of 1 per GPU. We train for 24 epochs using the Adam optimizer with an initial learning rate of 2×10^{-4} and applying a cosine annealing schedule. Data augmentation includes random scaling, cropping, color distortion, and Gridmask [4]. Each voxel receives a single label from the pre-generated occupancy ground truth.

4.3. Comparative Analysis

Since no occupancy-based approaches exist for UAV perception, we select two representative methods from the autonomous driving domain, BEVDet [14] and PanoOcc [34], and adapt them for our extended datasets. Because BEVDet [14] only produces 2D BEV predictions, we employ FlashOcc [43] to convert BEV features into 3D occupancy features, thereby allowing a consistent comparison with PanoOcc [34], which directly outputs occupancy predictions. As shown in Table 1, our experiments demonstrate that adopting occupancy features is advantageous for capturing both geometric and semantic information in 3D environments. Moreover, multi-UAV collaboration yields further gains in perception accuracy due to the widened coverage and shared information.

To highlight the advantages of our collaborative strategy, we implement collaborative approaches by integrating the Where2comm [12] module into BEVDet [14] and PanoOcc [34] for multi-UAV occupancy prediction. Experimental results demonstrate that our collaborative strategy achieves superior performance with significantly reduced communication overhead. Specifically, our method requires only 0.23 MB per transmission, while BEVDet

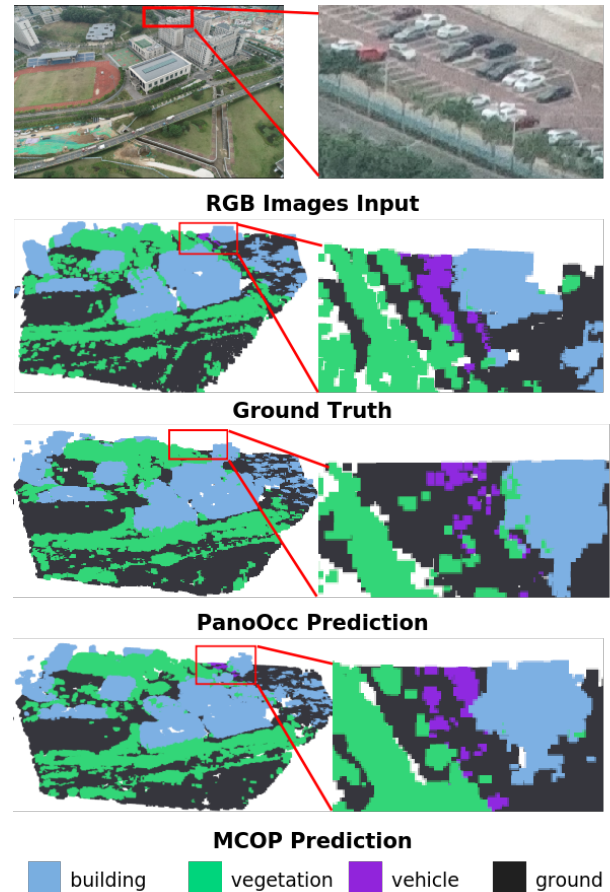


Figure 4. **Visualization results on GauUScene-Occ.** MCOP achieves better perception for distant and occluded objects.

Compress Method	Feature Dimensions	CV(MB)↓	mIoU↑	Accuracy Loss↓
-	3D	76.56	47.17	-
Avg.	2D	1.19	30.24	16.93%
Conv.	2D	1.19	41.52	5.64%
AAR	2D	1.19	46.42	0.75%

Table 2. **Different Feature Compression Strategies** Avg. represents the weighted average of features along the z-axis. Conv. refers to using a 3D convolution to convert 3D features into 2D features. Accuracy Loss represents the mIoU reduction ratio caused by compressing the features from 3D to 2D.

Connected Strategy	Quality Threshold	CV(MB)↓	mIoU↑	Accuracy Loss↓
Fully connected	-	1.19	46.42	-
Partially Connected	0.6	0.47	45.78	0.64%
Partially Connected	0.7	0.35	45.81	0.60%
Partially Connected	0.8	0.23	46.41	0.01%
Partially Connected	0.9	0.11	44.97	1.45%

Table 3. **Different Quality Score Threshold in DMPG**. All the above methods utilize the AAR module, and both Fully Connected and Partially Connected are based on 2D features.

and PanoOcc demand 17.50 MB and 19.14 MB, respectively. Furthermore, on the Air-to-Pred-Occ [35] dataset, our approach surpasses BEVDet by 35.12 mIoU points and PanoOcc by 4.45 points. These results confirm that compressing 3D occupancy features into altitude-aware 2D representations for transmission is more effective and efficient than direct 2D BEV transmission, balancing perceptual accuracy and communication efficiency effectively. Figure 4 shows sample results on GauUScene-Occ.

4.4. Ablation study

We assess the effectiveness of our modules by removing components under the same settings. Since our main experiments already show the advantage of generating and compressing 3D occupancy features over transmitting 2D BEV features, we focus on Altitude-Aware Reduction (AAR) and Dual-Mask Perceptual Guidance (DMPG), and also vary the number of UAVs.

Effectiveness of Altitude-Aware Reduction. In Table 2, we compare our proposed Altitude-Aware Reduction with two simpler compression methods: a weighted average along the z-axis and 3D convolution for compressing 3D features into 2D. All comparisons are conducted without the DMPG module. The comparison is performed by evaluating the prediction performance of the compressed features against the uncompressed ones. The uncompressed features retain the full 3D occupancy representation, resulting in a per-transmission communication cost of 76.56 MB. Our

Dataset	UAV Nums (mIoU ↑)				
	1	2	3	4	5
UAV3D	43.48	46.91	47.04	47.33	47.89
Air-Co-Pred	40.82	45.89	46.23	46.41	-
GauUScene	40.43	42.11	42.47	42.92	-

Table 4. **Impact of UAV quantity changes**. UAV num = 1 indicates no collaborative perception.

method reduces accuracy by only 0.75% at the same compression ratio, which is significantly better than the other two methods. This demonstrates that the introduced altitude encoding effectively preserves altitude information during compression, enabling the restoration of more comprehensive geometric and semantic details in feature fusion.

Effectiveness of Dual-Mask Perceptual Guidance. We also test different quality score thresholds, which govern whether UAVs transmit features based on perceived quality. A higher quality score threshold filters out regions of interest, thereby reducing the amount of data transmitted but risking insufficient information for accurate occupancy predictions. Conversely, a lower quality score threshold, while resulting in more regions being transmitted, can paradoxically lead to a decline in perception quality due to the inclusion of excessive irrelevant information. To evaluate the effect of different quality score threshold settings, we compare the mIoU accuracy loss between Partially Connected and Fully Connected modes. In this context, Partially Connected refers to transmitting only a portion of the features, while Fully Connected involves transmitting the complete set of features. Our baseline for comparison is the transmission of fully compressed features after applying the AAR module. We then assess various quality score threshold settings by comparing the mIoU accuracy loss between Partially Connected and Fully Connected modes. Table 3 shows that a 0.8 threshold achieves the best balance.

Impact of UAV quantity changes. Table 4 indicates that, under the same scenario, reducing the number of UAVs results in less than a 1% drop in perception accuracy, which remains higher than without collaborative perception. This demonstrates the robustness of our method to variations in UAV numbers.

5. Conclusion

This work proposes a multi-UAV collaborative occupancy prediction framework that addresses key limitations of existing BEV-based methods. Our framework effectively captures comprehensive geometric and semantic scene information while significantly reducing communication overhead. Extensive experiments on both virtual and real-world datasets demonstrate that our approach outperforms non-collaborative and alternative collaborative strategies, while requiring notably lower bandwidth.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. U21B2042, No. 62320106010). The authors would also like to thank the Key Laboratory of Target Cognition and Application Technology (TCAT), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100190, China, for their valuable support. In particular, we are grateful to Zhirui Wang and Peirui Cheng for their insightful discussions and assistance throughout the course of this research.

References

- [1] Saeed H. Alsamhi, Ou Ma, Mohammad Samar Ansari, and Faris A. Almalki. Survey on collaborative smart drones and internet of things for improving smartness of smart cities. *IEEE Access*, 7:128125–128152, 2019. 1, 2
- [2] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 1852–1864, 2020. 2
- [3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 3
- [4] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 7
- [5] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM MM*, pages 88–100, 2019. 3
- [6] Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *CVPR*, pages 17252–17262, 2022. 1, 2
- [7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 6
- [8] 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. 6
- [9] Chenyou Fan, Junjie Hu, and Jianwei Huang. Few-shot multi-agent perception with ranking-based feature learning. *PAMI*, 2023. 1
- [10] Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine*, 15(6):131–151, 2023. 2
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. 3, 4
- [12] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *NIPS*, 35: 4874–4886, 2022. 3, 7
- [13] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *CVPR*, pages 9243–9252, 2023. 3
- [14] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 7
- [15] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 3
- [16] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *NIPS*, 34:29541–29552, 2021. 3
- [17] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, pages 9087–9098, 2023. 3
- [18] Zhiqi Li, Wenhao Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18. Springer, 2022. 4
- [19] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 3
- [20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3, 6
- [21] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *CVPR*, pages 4106–4115, 2020. 3
- [22] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *ICRA*, pages 6876–6883. IEEE, 2020. 3
- [23] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *CVPR*, pages 19936–19945, 2024. 3
- [24] Thang Vu Myeongjin, Kim Jung-Hee, Jeong Seokwoo, and Seong-Gyun. Milo: Multi-task learning with localization ambiguity suppression for occupancy prediction cvpr 2023 occupancy challenge report. *arXiv preprint arXiv:2306.11414*, 2023. 3
- [25] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *ICRA*, pages 12404–12411. IEEE, 2024. 3
- [26] Donghao Qiao and Farhana Zulkernine. Adaptive feature fusion for cooperative perception using lidar point clouds. In *Winter CVPR*, pages 1186–1195, 2023. 2

- [27] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Aircsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018. 6
- [28] Sathira Silva, Savindu Bhashitha Wannigama, Roshan Ragel, and Gihan Jayatilaka. S2tpvformer: Spatio-temporal tri-perspective view for temporally coherent 3d semantic occupancy prediction. *arXiv e-prints*, page arXiv:2401.XXXX, 2024. 3
- [29] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *CVPR*, pages 17996–18006, 2024. 2, 3
- [30] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *NIPS*, 36, 2024. 2, 6
- [31] Nicholas Vadivelu, Mengye Ren, James Tu, Jingkang Wang, and Raquel Urtasun. Learning to communicate and correct pose errors. In *Conference on Robot Learning*, pages 1195–1210. PMLR, 2021. 3
- [32] Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *AAAI*, pages 5599–5606, 2024. 1
- [33] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *ECCV*, pages 605–621. Springer, 2020. 3
- [34] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *CVPR*, pages 17158–17168, 2024. 2, 3, 4, 7
- [35] Zhechao Wang, Peirui Cheng, Mingxin Chen, Pengju Tian, Zhirui Wang, Xinming Li, Xue Yang, and Xian Sun. Drones help drones: A collaborative framework for multi-drone object trajectory prediction and beyond. *ArXiv*, abs/2405.14674, 2024. 2, 6, 8
- [36] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *CVPR*, pages 21729–21740, 2023. 3
- [37] Xin Wu, Wei Li, Danfeng Hong, Ran Tao, and Qian Du. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):91–124, 2021. 3
- [38] Zhuoyuan Wu, Yuping Wang, Hengbo Ma, Zhaowei Li, Hang Qiu, and Jiachen Li. Cmp: Cooperative motion prediction with multi-agent communication. *arXiv preprint arXiv:2403.17916*, 2024. 1
- [39] Butian Xiong, Nanjun Zheng, Junhua Liu, and Zhen Li. Gauu-scene v2: Assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf, 2024. 2, 6
- [40] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022. 3
- [41] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *ECCV*, pages 107–124. Springer, 2022. 3
- [42] Hui Ye, Raj Sunderraman, and Shihao Ji. Uav3d: A large-scale 3d perception benchmark for unmanned aerial vehicles. In *The 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2, 6
- [43] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 3, 7
- [44] Haotian Zhang, Gaoang Wang, Zhichao Lei, and Jenq-Neng Hwang. Eye in the sky: Drone-based object tracking and 3d localization. In *ACM MM*, pages 899–907, 2019. 1
- [45] Yanan Zhang, Jinqing Zhang, Zengran Wang, Junhao Xu, and Di Huang. Vision-based 3d occupancy prediction in autonomous driving: a review and outlook. *arXiv preprint arXiv:2405.02595*, 2024. 3
- [46] Hao Zheng, Zhanlei Yang, Wenju Liu, Jizhong Liang, and Yanpeng Li. Improving deep neural networks using softplus units. In *IJCNN*, pages 1–4. IEEE, 2015. 7
- [47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. Last accessed: 17.11.2023. 4