VVSim: A Large-Scale Aerial-Ground Dataset and Benchmark for Cooperative Perception

Zengle Zhu[1*], Zhen Li[1*], Tianyi Huai[1], Tianshun Li[1], Zihang Xu[1], Liuqing Yang[1,2], Rongqing Zhang[1†], and Xinhu Zheng[1,2†]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology

{zzhu622, zli619, thuai231, tli449, zxu802}@connect.hkust-gz.edu.cn, lqyang@ust.hk, {rongqingz, xinhuzheng}@hkust-gz.edu.cn

Abstract. The increasing diversity and density of traffic participants have made modern driving environments highly complex, posing great challenges to reliable perception in autonomous driving. Single-vehicle perception is fundamentally limited by sensor range and occlusions, while vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) cooperation only partially mitigate these issues due to their fixed or ground-level viewpoints. In contrast, aerial platforms offer global and flexible sensing perspectives that can dynamically adjust their positions to provide broader spatial coverage and reduce blind zones. These advantages motivate the exploration of aerial-ground cooperative perception (AGCP). However, progress in this area is hindered by the lack of public datasets and standardized benchmarks. To bridge this gap, we introduce **VVSim**, a large-scale dataset for AGCP that provides synchronized multimodal data and state information from both vehicles and unmanned aerial vehicles (UAVs). VVSim consists of about 61k fully annotated frames spanning diverse interaction scenarios (e.g., cut-in, lane change), 5 weather conditions (sunny, foggy, rainy, cloudy, snowy), and 11 scenes (e.g., city, town, university, highway, mountain). In total, it comprises 255k LiDAR sweeps and 3.5M images (RGB, semantic segmentation, and depth) collected from vehicles and UAVs, along with detailed annotations for 2D/3D bounding boxes, object trajectories, and agent states. To support unified evaluation, we further propose **VVFormer**, a novel architecture that integrates multi-agent and multimodal features from UAVs and vehicles. Extensive experiments on VVSim demonstrate that VVFormer achieves superior performance, significantly outperforming strong V2V and V2I cooperative baselines on 3D perception tasks. The code is available at `https://github.com/LOTEAT/vvdetection3d`.

Keywords: Aerial-Ground Cooperative Perception · Multi-agent Multimodal Fusion · Datasets and Benchmark

[*] Zengle Zhu and Zhen Li contributed equally to this paper.
[†] Corresponding authors: Rongqing Zhang and Xinhu Zheng.

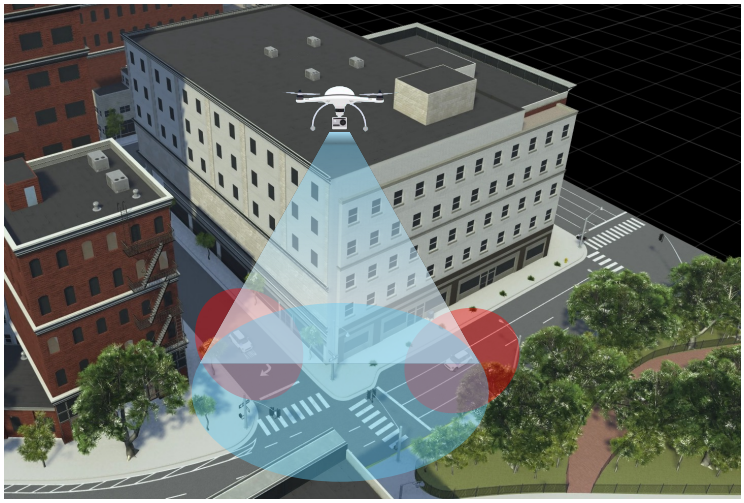

Fig. 1: Overview of the aerial-ground cooperative perception scenario. The aerial platform captures global context from a bird's-eye view, while ground vehicles perceive detailed local features. Their cooperation enhances perception coverage and mitigates occlusions.

1 Introduction

In recent years, technologies for intelligent driving have advanced rapidly, significantly improving autonomous driving systems [18]. These systems increasingly depend on accurate and real-time perception of the environment to detect nearby vehicles, pedestrians, and other traffic participants. It is critical for making safe and timely driving decisions [21]. Consequently, enhancing perception capabilities has become a central focus of autonomous driving research [7, 10, 12, 13, 17].

Relying on a single vehicle for perception often leads to blind spots. Onboard sensors have limited range and their view can be blocked by other vehicles or obstacles [30]. As a result, critical objects may be missed, reducing the reliability and safety of autonomous driving. To overcome these constraints, cooperative perception has emerged as a key paradigm, where multiple agents share information [6,11,30–33]. Among them, vehicle-to-vehicle (V2V) cooperation allows connected autonomous vehicles (CAVs) to complement sensing fields and mitigate occlusion. However, its performance is highly dependent on the spatial distribution of CAVs, limiting its widespread application. Vehicle-to-infrastructure (V2I) perception leverages roadside units (RSUs) to further extend the sensing range and alleviate occlusions. However, the fixed viewpoints of RSUs make it difficult to adapt to complex and dynamically changing traffic environments [23]. Moreover, the high deployment and maintenance costs hinder its large-scale adoption.

Compared with RSUs, unmanned aerial vehicles (UAVs) can flexibly adjust their altitude and position. As illustrated in Fig. 1, UAVs provide adaptive bird's-eye-view (BEV) perspectives that capture occluded areas and long-range

contexts inaccessible to ground sensors. In addition, UAVs offer wider coverage and lower deployment costs, enabling dynamic perception in highly dynamic and complex traffic scenes. Therefore, aerial-ground cooperative perception (AGCP) offers a promising paradigm to overcome the limitations of ground-based perception.

Despite the tremendous potential of aerial-ground cooperative perception, research in this area still faces several critical challenges. First, there is a lack of large-scale datasets that capture diverse environments and rich corner-case scenarios (e.g., severe occlusions, dense traffic interactions, and unusual vehicle behaviors), which are crucial for evaluating cooperative perception systems in safety-critical situations. Second, existing cooperative perception methods are typically designed for multiple agents with similar viewpoints and sensing configurations, making it difficult to model the heterogeneous nature of aerial-ground collaboration.

To address these gaps, we introduce VVSim, a large-scale aerial-ground cooperative perception dataset, together with VVFormer, a benchmark model specifically designed for evaluating aerial-ground cooperative perception.

The main contributions of this work are summarized as follows:

- We build **VVSim**, a large-scale aerial-ground cooperative perception dataset with time-synchronized and calibrated multimodal data collected from UAVs and vehicles. It features heterogeneous intelligent agents and diverse traffic participants, filling the gap in current datasets for aerial-ground cooperation across diverse environments with complex, high-density interactions.
- VVSim provides rich multimodal data and diverse driving conditions, including 61k fully annotated frames covering 19 interaction scenarios (e.g., cut-in, lane change, hard brake), 5 weather conditions, and 11 representative scenes (e.g., city, university, highway). In total, it contains 255k LiDAR sweeps and 3.5M images (RGB, semantic segmentation, depth) from both vehicles and UAVs. UAVs operate in hovering, cyclic, and escort modes at multiple altitudes, offering varied aerial viewpoints.
- We propose **VVFormer**, a benchmark model for aerial-ground cooperative perception. To the best of our knowledge, VVFormer is the first framework that performs cross-domain and multimodal cooperative perception between heterogeneous aerial and ground agents. Extensive experiments demonstrate that VVFormer achieves state-of-the-art performance and significantly outperforms strong ground-only cooperative baselines.

2 Related Works

2.1 Cooperative Perception

Cooperative perception has emerged as a promising approach to enhance environmental awareness for autonomous driving by enabling information sharing among multiple agents. Early studies primarily focus on V2V cooperation, where nearby vehicles exchange sensory features or object-level information to alleviate

occlusions and extend perception range [1,3,4,14,24]. These methods significantly improve detection accuracy but remain constrained by the limited local perspectives of ground vehicles. To further improve coverage and reliability, V2I perception frameworks incorporate RSUs equipped with LiDAR or cameras [27,30,33]. By integrating fixed sensing infrastructures, these approaches [15,28,34] achieve a broader field of view, thereby improving detection accuracy. However, fixed viewpoints limit adaptability in complex mixed-traffic scenarios, where dynamic participants and occlusions frequently occur. In addition, the high installation and maintenance costs of such infrastructures hinder scalability in large-scale or rapidly changing environments. Moreover, both V2V and V2I paradigms fundamentally rely on ground-level observations, leaving the perception of occluded or distant regions incomplete.

Despite progress in V2V and V2I perception, their effectiveness remains constrained by ground-level sensing and static viewpoints. V2V systems suffer from inconsistent coverage due to dynamic vehicle distributions, whereas V2I frameworks rely on costly and fixed infrastructure with limited adaptability. These limitations constrain the ability of agents to achieve comprehensive scene understanding, highlighting the need for a more flexible and globally aware cooperative perception paradigm.

UAVs provide flexible, elevated viewpoints that extend perception beyond the limits of ground sensors [5, 19]. They can adjust positions and altitudes in real time, offering wider coverage and better visibility in dynamic traffic scenes. Compared with RSUs, UAVs are easier to deploy and adapt quickly to changing environments. Leveraging these advantages, AGCP has emerged as a promising paradigm, offering complementary views that enhance detection performance and improve robustness under occlusion, long-range perception, and adverse weather [23, 26].

2.2 Cooperative Perception Dataset

Recent years have seen rapid growth in collaborative perception datasets for autonomous driving. OPV2V [31], DAIR-V2X [32], and V2X-Seq [33] laid the foundation for V2V and V2I cooperation. OPV2V provided synthetic multi-view data, while DAIR-V2X and V2X-Seq offered large-scale real-world and sequential multimodal data, respectively. V2V4Real further advanced real-traffic cooperative perception, yet all these datasets remain limited to ground-based sensors and cannot fully handle occlusion or blind-zone challenges. To address these limitations, several aerial-ground datasets have been introduced. CoPerception-UAV [8] and UAV3D [22] explore UAV-UAV perception in simulation. CoPeD presents a Vehicle-UAV dataset with real platforms, but UAVs fly at low altitudes (2-10 m) and only provide monocular data. V2U-COO [25] and Griffin [23] target Vehicle-UAV collaboration but contain limited interaction types and flight modes.

Existing aerial-ground datasets typically lack multi-mode UAV operations and multi-UAV collaboration, which limits their ability to capture the severe viewpoint variations inherent in aerial maneuvers. Furthermore, the number

of interacting UAVs is often limited, restricting the exploration of large-scale multi-agent cooperation in complex traffic environments. More importantly, autonomous driving systems must operate reliably under rare and safety-critical situations. Perception failures frequently arise in corner cases, such as severe occlusions, dense traffic interactions, unusual vehicle behaviors, and complex environmental conditions. However, existing cooperative perception datasets contain only limited coverage of such corner cases, making systematic evaluation under challenging scenarios difficult.

In contrast, our dataset provides a large-scale aerial-ground environment with rich corner-case scenarios for autonomous driving perception. It includes diverse scenes and interaction patterns, while UAVs operate under multiple flight modes and collaborate within a multi-agent framework. This design provides a realistic benchmark for studying cooperative perception in complex and safety-critical traffic situations.

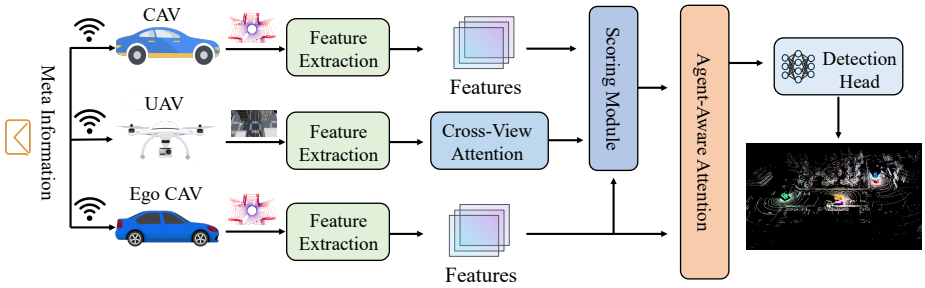

Fig. 2: Overall architecture of VVFormer. It consists of four key components: metadata sharing, feature extraction, cross-view attention, and agent-aware attention.

3 Methodology

In this section, we introduce VVFormer, an AGCP framework designed to address the lack of dedicated solutions for aerial-ground collaboration. Our method effectively fuses complementary information from UAVs and ground vehicles to achieve accurate and robust 3D scene understanding. Unlike traditional cooperative perception approaches that operate within a single domain, VVFormer unifies multimodal data from heterogeneous aerial and ground agents, providing a comprehensive solution for AGCP. In the following, we describe the overall architecture and key components of our method in detail.

3.1 Main Architecture Design

The proposed VVFormer is designed to achieve comprehensive cooperative perception through multi-agent and multimodal fusion between UAVs and vehi-

cles. Specificall, to effectively leverage complementary perspectives and sensing modalities, VVFormer adopts a unified architecture that integrates both aerial and ground agents. As illustrated in Fig. 2, the framework is composed of four main components: metadata sharing, feature extraction, cross-view attention, and agent-aware attention. These modules jointly enable spatially consistent feature alignment and interaction across heterogeneous agents, facilitating accurate and robust scene understanding under diverse driving and flight conditions.

3.2 Metadata Sharing

During the early stage of collaboration, every agent $i \in \{1, 2, \ldots, N\}$ broadcasts essential metadata that describe its state. The shared information includes the ego pose in the global frame, timestamps, and the intrinsic and extrinsic parameters of sensors. These metadata are exchanged among all agents, including both vehicles and UAVs, providing the basic system states necessary for cooperative perception.

3.3 Feature Extraction

For each ground vehicle, the LiDAR point cloud is divided into $H_l \times W_l$ pillars along the BEV plane, with each pillar corresponding to a fixed region on the ground. PointPillars [10] converts points within each pillar to a feature vector, producing a BEV feature map $F_i^l \in \mathbb{R}^{H_l \times W_l \times C}$, where C denotes the number of feature channels. Each spatial location in F_i^l corresponds to a pillar in the BEV grid.

For UAVs, RGB images are processed by a Swin Transformer [16] to produce a feature map $F_i^c \in \mathbb{R}^{H_c \times W_c \times C}$, where the channel dimension C aligns with that of the LiDAR feature map, enabling consistent feature fusion across modalities.

3.4 Cross-View Attention

To achieve spatial alignment between UAV image features and ground LiDAR features, we design a ***cross-view attention*** (CVA) module that transforms UAV-centric image features into the vehicle-centric BEV coordinate system, which is shown in Fig. 3. This module enables direct fusion between heterogeneous modalities while preserving geometric consistency across different viewpoints.

Voxel Projection. To capture the vertical structure of the scene and enable interaction with aerial observations, we extend the BEV pillars from ground vehicles along the height dimension. We first construct a 3D voxel grid, resulting in a grid of size $H_l \times W_l \times D_{\text{vox}}$. Each voxel center $p_v = (x_v, y_v, z_v, 1)^\top$ is represented in global coordinates. To obtain the corresponding location in the UAV image, we transform each voxel from global coordinates to the camera coordinate system using the extrinsic matrix $T_{\text{ext}} \in \mathbb{R}^{4\times4}$ of the UAV

$$p_c = T_{\text{ext}}\, p_v = (x_c, y_c, z_c, 1)^\top. \tag{1}$$

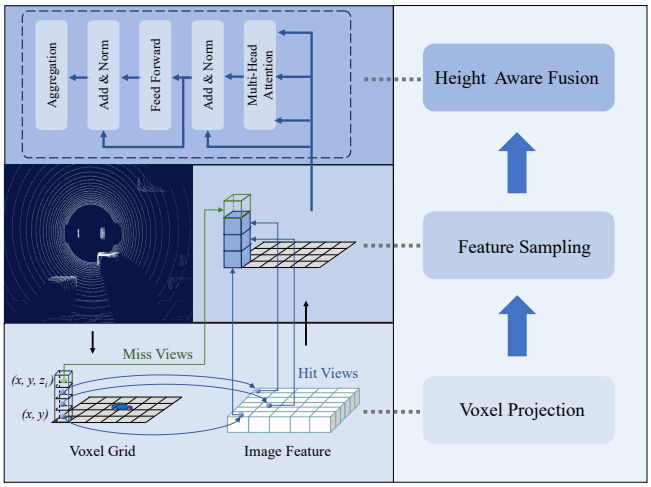

Fig. 3: The architecture of cross-view attention.

We then project 3D camera coordinates to homogeneous image coordinates using the intrinsic matrix $T_{\text{int}} \in \mathbb{R}^{3\times3}$

$$u_{\text{hom}} = T_{\text{int}} \left[x_c, y_c, z_c\right]^\top = (u_h, v_h, w_h)^\top. \tag{2}$$

Then we obtain the pixel coordinates by normalization

$$u = \frac{u_h}{w_h}, \quad v = \frac{v_h}{w_h}. \tag{3}$$

A validity mask $\mathcal{M}$ is applied to exclude voxels outside the image boundaries or behind the camera:

$$\mathcal{M}(p) = \begin{cases} & 0 \le u < W_{\text{img}}, \\ 1, & 0 \le v < H_{\text{img}}, \\ & z_c > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

H_{img} and W_{img} denote the height and width of the UAV image, and z_c is the depth in the camera frame.

The projected coordinates are then scaled according to the backbone downsampling ratio s:

$$(u', v') = \left\lfloor \frac{u}{s} \right\rfloor, \left\lfloor \frac{v}{s} \right\rfloor \tag{5}$$

establishing a direct correspondence between each voxel and the downsampled feature map.

Feature Sampling. After constructing the voxel grid and projecting it onto the UAV image features, we sample image features to build a voxel-aligned BEV representation. For each voxel at position (x_v, y_v, z_v) with a validity mask

$\mathcal{M}(x_v, y_v, z_v) = 1$, the corresponding pixel in the UAV feature map F_i^c is sampled. Voxels with $\mathcal{M}(x_v, y_v, z_v) = 0$ are filled with zeros.

Formally, let the projected coordinates on the downsampled feature map be (u', v'), then the sampled feature F_v for voxel (x_v, y_v, z_v) is

$$F_v = \begin{cases} F_i^c(u', v'), & \mathcal{M}(x_v, y_v, z_v) = 1, \\ 0, & \mathcal{M}(x_v, y_v, z_v) = 0. \end{cases} \tag{6}$$

Repeating this process for all voxels along the height dimension D_{vox}, we obtain a voxel-aligned BEV feature map

$$F_i^{c2v} \in \mathbb{R}^{D_{\text{vox}} \times H_l \times W_l \times C} \tag{7}$$

where each slice along D_{vox} corresponds to features sampled from a different height level. This feature map is now spatially aligned with the LiDAR BEV features F_i^l, enabling effective cross-modal fusion in subsequent modules.

Height-Aware Fusion. After obtaining the voxel-aligned image features from the projected feature sampling module, we perform ***height-aware fusion*** to aggregate information across different heights. Specifically, each voxel-aligned feature $F_i^{c2v} \in \mathbb{R}^{D_{\text{vox}} \times H_l \times W_l \times C}$ is indexed by its height d, and spatial coordinates (h, w) in the BEV plane. To provide explicit geometric cues along the height axis, we add a height positional encoding to each voxel slice:

$$\tilde{F}_i^{c2v}(d, h, w) = F_i^{c2v}(d, h, w) + P_{\text{height}}(d). \tag{8}$$

$P_{\text{height}}(d) \in \mathbb{R}^C$ is the positional encoding for the d-th height level, defined as

$$P_{\text{height}}(d) = \begin{cases} \sin\Big(\dfrac{d}{10000^{2k/C}}\Big), & \text{if } d = 2k, \\ \cos\Big(\dfrac{d}{10000^{2k/C}}\Big), & \text{if } d = 2k+1 \end{cases} \tag{9}$$

where $k \in [0, \lfloor C/2 \rfloor)$.

We then apply height-wise self-attention for each spatial location (h, w). The attention weights along the height dimension are computed as

$$A(d, d') = \frac{\exp\left(Q_{h,w,d} \cdot K_{h,w,d'} / \sqrt{C}\right)}{\sum_{d''=1}^{D_{\text{vox}}} \exp\left(Q_{h,w,d} \cdot K_{h,w,d''} / \sqrt{C}\right)} \tag{10}$$

where $Q, K \in \mathbb{R}^{D_{\text{vox}} \times C}$ are linear projections of $\tilde{F}_i^{c2v}$ along the height dimension, and the fused feature is obtained by

$$F_i^{\text{fused}}(h, w) = \sum_{d=1}^{D_{\text{vox}}} A(d, :)\, V_{h,w,d}, \tag{11}$$

with V being the value projection of $\tilde{F}_i^{c2v}$. The resulting BEV feature map $F_i^{\text{fused}} \in \mathbb{R}^{C \times H_l \times W_l}$ now incorporates height-aware context.

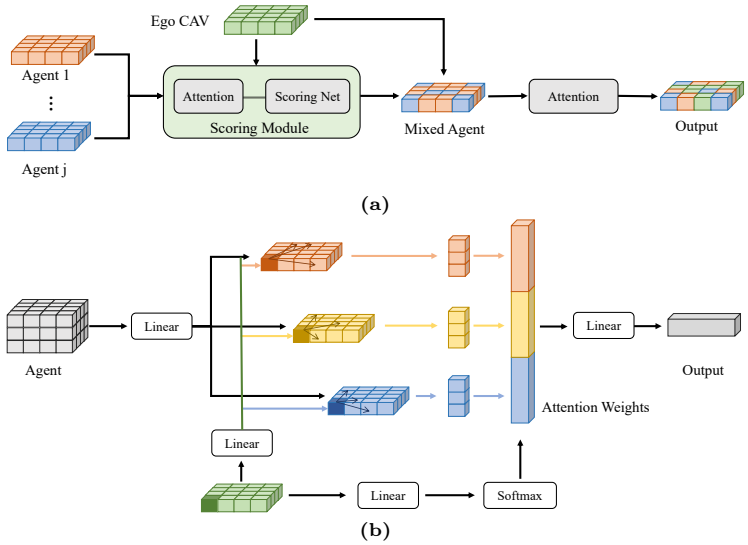

Fig. 4: Agent-aware attention for AGCP. (a) Overview of the agent-aware attention mechanism. (b) Illustration of the deformable attention module used within agent-aware attention.

3.5 Agent-Aware Attention

To enable collaborative perception across heterogeneous agents such as connected vehicles and UAVs, we design an ***agent-aware attention*** (AAA) module that adaptively selects and integrates the most informative features from multiple viewpoints, as illustrated in Fig. 4. The core idea is to assign an importance score to each agent at every BEV grid location, and then refine the aggregated representation through a spatial cross attention.

Agent Cross Attention Scoring. Given BEV features $\{F_i \in \mathbb{R}^{H_l \times W_l \times C}\}_{i=1}^{N}$ from N collaborating agents, our goal is to adaptively identify which agent provides the most informative observation for each spatial cell in the BEV grid. We first take the ego agent as the query source and the other agents as key-value sources. Each feature map F_i is flattened into $X_i \in \mathbb{R}^{S \times C}$, where $S = H_l \times W_l$ denotes the number of BEV grids.

For each non-ego agent j, we perform deformable cross-attention operation between the ego query X_{ego}, reference points R, and value X_j:

$$X_j^{\text{out}} = \text{DCA}(X_{\text{ego}}, R, X_j) \tag{12}$$

where $\text{DCA}(\cdot)$ denotes the deformable cross-attention, formulated as:

$$X_j^{\text{out}}(q) = \sum_{m=1}^{M} W_m \sum_{k=1}^{K} A_{mqk}\, W_v\, X_j(p_q + \Delta p_{mqk}) \tag{13}$$

where q indexes the query position, M is the number of attention heads, and K is the number of sampling points for each head. Δp_{mqk} represents the learnable offsets, A_{mqk} are the normalized attention weights, W_v denotes the value projection matrix, W_m represents the weights of attention head.

Next, each cross-attention output X_j^{out} is fed into a lightweight *ScoringNet* to estimate an importance score map $S_j \in \mathbb{R}^{S\times 1}$

$$S_j = \text{ScoringNet}(X_j^{\text{out}}) \tag{14}$$

where each score $S_j(q)$ measures the informativeness of agent j at BEV position q. For each spatial cell, we then select the feature from the agent with the highest score

$$X^*(q) = X^{\text{out}}_{j^*(q)}(q), \quad j^*(q) = \arg\max_j S_j(q). \tag{15}$$

This process adaptively aggregates the most informative agent features across the BEV grid.

Spatial Cross Attention. After obtaining the agent-aware BEV feature X^*, we further enhance spatial consistency by applying a spatial cross attention between the ego BEV query X_{ego} and the aggregated value feature X^*.

For each query position i, we assign a normalized reference point $r_i \in [0,1]^2$ corresponding to its spatial coordinate in the BEV plane. These reference points serve as anchors that guide each query to attend to spatially relevant regions. To capture fine-grained spatial interactions, multiple learnable offsets $\Delta r_{h,l}$ are applied around each reference point, allowing the query to adaptively sample contextual features from nearby locations.

Formally, the spatial cross attention is expressed as

$$X^{\text{out}}(i) = \sum_{h=1}^{H} W_h \sum_{l=1}^{L} A_{hil}\, W_v\, X^*\big(r_i + \Delta r_{hil}\big) \tag{16}$$

where H is the number of attention heads, L is the number of sampling points per head, A_{hil} are normalized attention weights, W_v is a learnable projection matrix, and Δr_{hil} denote the learnable 2D sampling offsets. In practice, the reference points r_i are initialized on a uniform BEV grid and iteratively refined through the attention layers, ensuring that the attention field aligns with meaningful spatial structures.

This spatial cross attention enables each BEV query to aggregate information from spatially adjacent, resulting in a feature representation that maintains both spatial continuity and inter-agent contextual awareness.

4 VVSim Dataset

To facilitate research on AGCP, we develop **VVSim**, a large-scale multimodal dataset providing synchronized sensor observations and state information from both ground vehicles and UAVs. VVSim is designed to capture the complementary viewpoints inherent in heterogeneous agents and to support a broad range of cooperative perception tasks.

4.1 Data Collection Setup

VVSim is built on top of the AirSim, which we extend with customized modules for multi-agent coordination and high-fidelity scene rendering. The platform supports various UAV flight modes (hovering, cyclic, escort), enabling diverse aerial-ground cooperative configurations.

Each agent is equipped with a set of sensors operating at 25 Hz. Ground vehicles mount:

- 4 RGB cameras,
- 4 semantic segmentation cameras,
- 4 depth cameras,
- 1 top-mounted LiDAR.

The cameras are placed at four viewpoints (front, left, right, rear) on the roof, providing full 360° visual coverage. Each UAV carries one downward-facing RGB camera and one semantic segmentation camera, mounted centrally beneath the drone body to capture wide-area overhead observations.

4.2 Dataset Composition

VVSim contains approximately **61k** fully annotated frames, each offering synchronized multimodal observations from all participating agents. In total, the dataset includes:

- **255k** LiDAR sweeps,
- **3.5M** images consisting of **1.2M RGB** images, **1.1M depth** images, and **1.2M semantic segmentation** images,
- **986k** 3D bounding box annotations across seven object categories: *car, pedestrian, truck, trailer, traffic cone, van, speed limit sign*,
- **object trajectories** of traffic participants,
- **6-DoF agent states**, including position (x, y, z) and orientation (roll, pitch, yaw), along with velocity and control-related information.

All annotations follow physically consistent temporal alignment, supporting detection, tracking, prediction, and other cooperative tasks.

4.3 Scenes and Scenarios

VVSim covers various scenes and traffic conditions:

- **11 scenes** spanning city, town, university, highway, and other representative environments,
- **5 weather conditions**: sunny, foggy, rainy, cloudy, and snowy,
- **19 traffic interaction scenarios**, including cut-in, lane change, overtaking, merging, and other common interaction patterns.

Each scenario includes multiple ground vehicles and UAVs, forming complex multi-agent configurations and dense traffic behaviors suitable for testing aerial-ground cooperative perception algorithms.

5 Experiments

5.1 Experimental Settings

Dataset. We construct a multi-agent collaborative perception dataset, comprising sequences captured by both connected vehicles and UAVs. In total, the dataset contains 61625 frames. For evaluation purposes, the dataset is divided into training, validation, and testing sets, containing 38000, 5125, and 18500 frames, respectively. This split ensures a balanced distribution of diverse traffic scenarios and environmental conditions across different sets.

Setup. The point cloud range is set to $x, y \in [-50, 50]$ meters. The BEV representation is discretized with a grid resolution of 0.5 meters. The model is trained for 24 epochs using an AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 0.01.

Task and Metrics. Our framework supports multiple cooperative perception tasks, such as object detection and tracking. In this paper, we focus on aerial-ground cooperative 3D object detection, where the goal is to jointly perceive the surrounding environment using both aerial and ground agents. The evaluation follows the standard metrics defined in the nuScenes benchmark [2], including mean Average Precision (mAP), mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and the overall nuScenes Detection Score (NDS).

Baselines. To better evaluate the performance of our model, we use the following baseline models.

- F-Cooper [3]: An early cooperative perception framework that uses sparse point cloud object detection.
- OPV2V [31]: A V2V cooperative perception system that applies attentive intermediate fusion.
- V2X-ViT [30]: A transformer-based V2X framework that uses heterogeneous multi-agent self-attention.
- CoBEVT [29]: A model that employs fused axial attention to integrate multi-agent features.
- SICP [20]: A lightweight dual-perception network that enables collaborative 3D detection.
- L4DR [9]: A weather-robust 3D detection model that enhances perception via foreground-aware denoising.

5.2 Quantitative Results

Table 1 shows a comprehensive comparison of proposed VVFormer with representative cooperative perception methods on the VVSim test set. Our approach achieves the best overall performance, with an mAP of 0.6490 and an NDS of 0.5268, clearly outperforming existing methods across most metrics. The substantial improvement in overall perception quality demonstrates the effectiveness

Table 1: Comparison with advanced methods on the VVSim test set.

Method	Publication	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓
F-Cooper [3]	SEC2019	0.5865	0.4773	0.3530	0.1144	0.6915	2.7622
V2XViT [30]	ECCV2022	0.6150	0.4847	0.3455	0.1061	0.7766	**2.6077**
CoBEVT [29]	CoRL2023	0.5777	0.4654	0.3681	0.1544	0.7124	2.9697
SICP [20]	IROS2024	0.6327	0.5144	0.2263	0.1055	0.6877	3.5578
L4DR [9]	AAAI2025	0.6420	0.4924	0.3595	0.1408	0.7862	2.9374
VVFormer (ours)	-	**0.6490**	**0.5268**	**0.2168**	**0.0954**	**0.6647**	3.2290

of our multi-agent fusion strategy in achieving consistent and reliable 3D understanding.

In terms of position and scale estimation, VVFormer attains the lowest mATE (0.2168) and mASE (0.0954), indicating its advancement in percepting precise bounding boxes and accurate object positions. The mAOE (0.6647) is also superior to other approaches, suggesting that our model provides more accurate orientation predictions.

It is worth noting that the mAVE (3.2290) is slightly higher than the best-performing baseline (F-Cooper). This can be attributed to the wide speed distribution of dynamic agents in the VVSim test set, which includes vehicles and diverse motion patterns spanning both low-speed urban maneuvers and high-speed transit conditions. Such a diverse motion spectrum increases the difficulty of temporal velocity estimation, particularly in distant or intermittently observed objects. Nonetheless, VVFormer still maintains competitive motion estimation capability while achieving a state-of-the-art (SOTA) performance across other major metrics. Overall, these results validate the effectiveness of our cooperative perception framework in enhancing scene understanding under complex real-world conditions.

5.3 Ablation Study

Table 2 reports the ablation results of each major modules in our framework on the VVSim dataset. Specifically, We analyze three critical modules: the Cross-View Attention (CVA) module, the Agent-Aware Attention (AAA) module, and the contribution of UAVs.

Effectiveness of UAVs. We can observe that introducing UAVs significantly improves the overall perception performance (mAP from 0.5294 to 0.6063 and NDS from 0.4550 to 0.4964). This clearly validates the effectiveness of aerial-ground cooperation. The performance improvement can be attributed to the complementary top-down viewpoints from UAVs. These viewpoints expand spatial coverage of the environment and alleviate occlusions that are prevalent in ground-only settings.

Effectiveness of Agent-Aware Attention. The AAA module plays a key role in feature interaction among cooperation agents. When AAA is removed, the network struggles to integrate the heterogeneous feature representations from

Table 2: The experimental ablation results of each module on VVSim test set.

CVA	AAA	UAVs	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓
✓	✓		0.5912	0.4912	0.2459	0.0991	0.6992	3.3559
✓		✓	0.6063	0.4964	0.2613	0.1089	0.6972	2.5410
	✓	✓	0.5294	0.4550	0.2964	0.1286	0.6720	3.2534

Note: CVA = Cross-View Attention; AAA = Agent-Aware Attention; UAVs indicates removing UAV perception data.

different agents, leading to significant performance degradation across all metrics. In contrast, incorporating AAA allows the model to learn agent-dependent feature weights, enabling more effective multi-agent fusion and thus improving overall perception performance.

Effectiveness of Cross-View Attention. The CVA module contributes to cross-view alignment between feature maps extracted from different sensing perspectives. CVA facilitates fine-grained alignment in the BEV space, ensuring that semantically corresponding regions from distinct views are properly matched. This leads to improved detection performance.

Overall, the ablation results confirm that both the aerial-ground cooperation and the attention-based alignment modules (AAA and CVA) are essential for achieving accurate and robust cooperative perception in complex multi-agent environments.

6 Conclusion

In this work, we present VVSim, a large-scale dataset designed to advance research in aerial-ground cooperative perception. VVSim provides synchronized multi-modal data from both UAVs and vehicles, covering various weather conditions, scenes, and interaction scenarios. Along with the dataset, we introduce VVFormer, a unified architecture that effectively fuses multi-agent and multi-modal features from aerial and ground platforms, serving as a strong baseline for cooperative perception tasks. Together, VVSim and VVFormer establish the comprehensive foundation for evaluating and developing aerial-ground collaborative perception systems.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant 62271351, U24A20252, 62373315 and U23A20339, in part by Guangzhou Municipal Science and Technology Project 2024D03J0008, and Guangdong Provincial Project 2023ZDZX1037 and 2023ZT10X009, in part by New Generation Artificial Intelligence-National Science and Technology Major Project No. 2025ZD0124202, in part by National Key Research and Development Program of China Grant 2024YFB4707603, and in part by Nansha Key Science and Technology Project 2023ZD006.

References

1. Arnold, E., Dianati, M., De Temple, R., Fallah, S.: Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. IEEE Transactions on Intelligent Transportation Systems **23**(3), 1852–1864 (2020)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
3. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 514–524. IEEE (2019)
4. Chiu, H.k., Hachiuma, R., Wang, C.Y., Smith, S.F., Wang, Y.C.F., Chen, M.H.: V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. arXiv preprint arXiv:2502.09980 (2025)
5. Fan, J., Fan, L., Ni, Q., Wang, J., Liu, Y., Li, R., Wang, Y., Wang, S.: Perception and planning of intelligent vehicles based on bev in extreme off-road scenarios. IEEE Transactions on Intelligent Vehicles **9**(4), 4568–4572 (2024)
6. Hao, R., Fan, S., Dai, Y., Zhang, Z., Li, C., Wang, Y., Yu, H., Yang, W., Yuan, J., Nie, Z.: Rcooper: A real-world large-scale dataset for roadside cooperative perception. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22347–22357 (2024)
7. Hu, H., Wang, F., Su, J., Wang, Y., Hu, L., Fang, W., Xu, J., Zhang, Z.: Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. arXiv preprint arXiv:2303.17895 (2023)
8. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
9. Huang, X., Xu, Z., Wu, H., Wang, J., Xia, Q., Xia, Y., Li, J., Gao, K., Wen, C., Wang, C.: L4dr: Lidar-4dradar fusion for weather-robust 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3806–3814 (2025)
10. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
11. Li, B., Xu, Z., Li, J., Liu, X., Fang, J., Li, X., Yu, H.: V2x-dg: Domain generalization for vehicle-to-everything cooperative perception. arXiv preprint arXiv:2503.15435 (2025)
12. Li, Z., Lan, S., Alvarez, J.M., Wu, Z.: Bevnext: Reviving dense bev frameworks for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20113–20123 (2024)
13. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022)
14. Lin, C., Tian, D., Duan, X., Zhou, J., Zhao, D., Cao, D.: V2vformer: Vehicle-to-vehicle cooperative perception with spatial-channel transformer. IEEE Transactions on Intelligent Vehicles **9**(2), 3384–3395 (2024)

15. Liu, S., Ding, Z., Fu, J., Li, H., Chen, S., Zhang, S., Zhou, X.: V2x-pc: Vehicle-to-everything collaborative perception via point cluster. arXiv preprint arXiv:2403.16635 (2024)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 2774–2781. IEEE (2023)
18. Lu, Y., Ma, H., Smart, E., Yu, H.: Real-time performance-focused localization techniques for autonomous vehicle: A review. IEEE Transactions on Intelligent Transportation Systems **23**(7), 6082–6100 (2021)
19. Mohsan, S.A.H., Othman, N.Q.H., Li, Y., Alsharif, M.H., Khan, M.A.: Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends. Intelligent service robotics **16**(1), 109–137 (2023)
20. Qu, D., Chen, Q., Bai, T., Lu, H., Fan, H., Zhang, H., Fu, S., Yang, Q.: Sicp: Simultaneous individual and cooperative perception for 3d object detection in connected and automated vehicles. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8905–8912. IEEE (2024)
21. Sun, C., Zhang, R., Lu, Y., Cui, Y., Deng, Z., Cao, D., Khajepour, A.: Toward ensuring safety for autonomous driving perception: Standardization progress, research advances, and perspectives. IEEE Transactions on Intelligent Transportation Systems **25**(5), 3286–3304 (2023)
22. Sunderraman, R., Ji, J.S., et al.: Uav3d: A large-scale 3d perception benchmark for unmanned aerial vehicles. Advances in Neural Information Processing Systems **37**, 55425–55442 (2024)
23. Wang, J., Cao, X., Zhong, J., Zhang, Y., Yu, H., He, L., Xu, S.: Griffin: Aerial-ground cooperative detection and tracking dataset and benchmark. arXiv preprint arXiv:2503.06983 (2025)
24. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: European conference on computer vision. pp. 605–621. Springer (2020)
25. Wang, Y., Wang, Z., Cheng, P., Tian, P., Yuan, Z., Tian, J., Wang, W., Zhao, L.: Uvcpnet: A uav-vehicle collaborative perception network for 3d object detection. IEEE Transactions on Geoscience and Remote Sensing (2025)
26. Wu, J., Zhang, N., Li, D., Bi, J., Han, G.: A context-aware feature fusion method for multi-uav cooperative air combat. IEEE Transactions on Intelligent Transportation Systems (2025)
27. Xiang, H., Zheng, Z., Xia, X., Xu, R., Gao, L., Zhou, Z., Han, X., Ji, X., Li, M., Meng, Z., et al.: V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In: European Conference on Computer Vision. pp. 455–470. Springer (2024)
28. Xu, R., Chen, C.J., Tu, Z., Yang, M.H.: V2x-vitv2: Improved vision transformers for vehicle-to-everything cooperative perception. IEEE transactions on pattern analysis and machine intelligence (2024)
29. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202 (2022)

30. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022)
31. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)
32. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., 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. pp. 21361–21370 (2022)
33. Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., 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. pp. 5486–5495 (2023)
34. Zhou, Z., Xiang, H., Zheng, Z., Zhao, S.Z., Lei, M., Zhang, Y., Cai, T., Liu, X., Liu, J., Bajji, M., et al.: V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25399–25409 (2025)