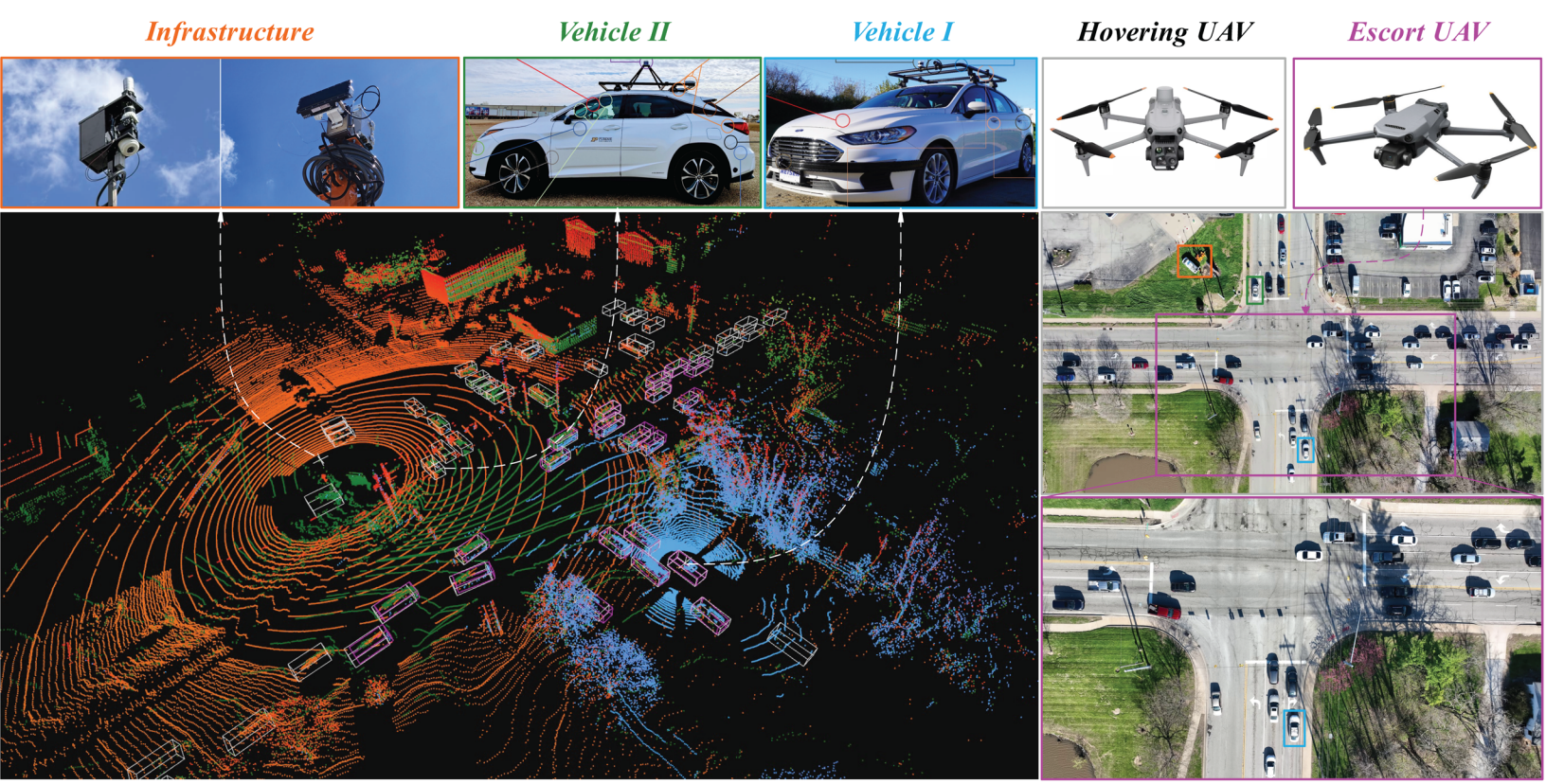
CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Final Report #109
September 2026



LiDAR Point Cloud (Vehicle & Infrastructure) & UAV Detections (Hovering and Escort)

UAV Camera (Hovering and Escort)

Cost Benefit Analysis of an Infrastructure-Vehicle Cooperative Perception System

Yiheng Feng
Ziran Wang
Andrew Tarko

PURDUE UNIVERSITY®

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Report No. 109 **August 2026**

Project Start Date: 06/01/2024

Project End Date: 05/31/2026

Cost Benefit Analysis of an Infrastructure-Vehicle Cooperative Perception System

by

Yiheng Feng, Associate Professor, Purdue University

Ziran Wang, Assistant Professor, Purdue University

Andrew Tarko, Professor, Purdue University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3551747105 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Feng, Y., Wang, Z., & Tarko, A. (2026). Cost Benefit Analysis of an Infrastructure-Vehicle Cooperative Perception System. Final Report. USDOT CCAT Project No. 109.

Contacts

For more information:

Dr. Yiheng Feng
HAMP G131
550 Stadium Mall Drive
West Lafayette, IN 47907
Phone: 765-496-5025
Email: feng333@purdue.edu

CCAT
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48152
umtri-ccat@umich.edu
(734) 763-2498

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Technical Report Documentation Page

1. Report No. Report 109	**2. Government Accession No.** Leave blank – not used	**3. Recipient's Catalog No.**
4. Title and Subtitle Cost Benefit Analysis of an Infrastructure-Vehicle Cooperative Perception System		**5. Report Date** August 2026 **6. Performing Organization Code** XXXX
7. Author(s) Feng, Yiheng, Ph.D., https://orcid.org/0000-0001-5656-3222 Wang, Ziran, Ph.D., https://orcid.org/ 0000-0003-2702-7150 Tarko, Andrew, Ph.D., https://orcid.org/0000-0001-9301-3270		**8. Performing Organization Report No.** **XXXX**
9. Performing Organization Name and Address Center for Connected and Automated Transportation University of Michigan Transportation Research Institute 2901 Baxter Road Ann Arbor, MI 48109.		**10. Work Unit No.** **11. Contract or Grant No.** Contract No. 69A3551747105
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Report (June 2024 – May 2026) **14. Sponsoring Agency Code** OST-R

15. Supplementary Notes
Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.

16. Abstract
In this project, we develop a real-world Air-Ground-Infrastructure cooperative perception dataset collected using two instrumented vehicles, one roadside sensing unit, and two camera-equipped unmanned aerial vehicles (UAVs). The UAVs operate in two complementary modes: a hovering UAV provides persistent coverage of the study area, while an escort UAV follows an instrumented vehicle to capture mobile aerial observations. This dataset enables controlled evaluation of vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-air (V2A), and full vehicle-to-infrastructure-to-air (V2I2A) cooperation. We benchmark representative early-, intermediate-, and late-fusion methods under conventional ground-agent configurations (V2V and V2I) and further evaluate the contribution of UAV-derived detections through late-stage fusion (V2A and V2I2A).

17. Key Words Cooperative Perception; Vehicle-infrastructure Cooperation; Air-ground Cooperation; Unmanned Aerial Vehicles; Dataset; V2X	**18. Distribution Statement** No restrictions.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 19	**22. Price** Leave blank – not used

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

UMTRI · The University of Akron · Central State University · Illinois Center for Transportation · University of Minnesota · Northwestern · Purdue University · Washtenaw Community College · Wisconsin – University of Wisconsin-Madison

Table of Contents

1. Introduction 2
2. System Development 3
   2.1 Sensor Specifications 3
   2.2 Data Collection 4
   2.3 UAV Detection and Tracking 5
   2.4 Coordinate System 6
   2.5 Data Annotation 7
3. Findings 7
4. Recommendations 11
5. Outputs, Outcomes, and Impacts 12

References 14

List of Figures

Figure 1 Demonstration of data acquisition systems. 4
Figure 2 Demonstration of UAV detections. 6
Figure 3 Demonstration of mapping products 7
Figure 4 Qualitative comparison of Early Fusion, HEAL, and Late Fusion under V2V, V2I, and V2X cooperative topologies. 11

List of Tables

Table 1 Sensor configuration of the air–ground–infrastructure data-collection system. 3
Table 2 Cooperative perception settings evaluated in AGI-Coop. Veh. I is the ego vehicle in all multi-agent configurations. 8
Table 3 Comparison of Fusion Strategies under Canonical V2V, V2I, and V2X Topologies on the AGI-COOP Dataset. The best and second-best results are highlighted in bold and underlined, respectively. 9
Table 4 Scalability of Ego-Centric Cooperative Perception under Different Vehicle-Infrastructure-Aerial Agent Configurations. The performance gains are calculated relative to the Ego-only baseline. 10

1. Introduction

Sensing and perception of the driving environment and traffic conditions are critical for autonomous driving and intelligent transportation system (ITS) applications. Video camera, radar, and LiDAR are commonly used sensors for perception of vehicles and the infrastructure. Perception sensors implemented only in vehicles or only at the roadside are insufficient for many applications due to their limited accuracy, range, coverage, and occlusion, which leads to insufficient accuracy of the results. A cooperative perception system that shares, processes, and fuses sensor data from multiple sources in real-time and is further leveraged with real-time communication technologies such as C-V2X or 5G. Such as system can greatly improve data collection efficiency, and it can provide a more complete picture of the road conditions and traffic states. Recent research has shown the effectiveness of cooperative perception that utilizes collected and shared traffic data from a limited number of perception sensors, to achieve a high augmented penetration rate [1][2]. However, existing studies are limited in three aspects: 1) most of the studies are conducted in a simulation environment with simulated sensor data [3] and/or simplified assumptions in detection models (e.g., perfect detection within a predefined range) [2]; 2) the data modality is limited only to infrastructure (e.g., [4]) or only to vehicles (e.g., [5]); and 3) almost all the existing vehicle-infrastructure cooperative perception datasets such as [4,5,6] require manual annotation of 3D labels, which is both time consuming and costly. Moreover, most recent studies only collected and shared data, and did not conduct further analysis on the benefit of the cooperative perception system on downstream applications and associated costs compared with single (or traditional) sensor settings.

To address these limitations, we develop a real-world Air-Ground-Infrastructure cooperative perception dataset (**AGI-Coop**), collected using two instrumented vehicles, one roadside sensing unit, and two camera-equipped unmanned aerial vehicles (UAVs). The UAVs operate in two complementary modes: a hovering UAV provides persistent coverage of the study area, while an escort UAV follows an instrumented vehicle to capture mobile aerial observations. AGI-Coop enables controlled evaluation of vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-air (V2A), and full vehicle-to-infrastructure-to-air (V2I2A) cooperation. We benchmark representative early-, intermediate-, and late-fusion methods under conventional ground-agent configurations (V2V and V2I) and further evaluate the contribution of UAV-derived detections through late-stage fusion (V2A and V2I2A). Furthermore, we develop a new auto-labeling method called DAA (Drones as Annotators), which augments onboard LiDAR with an independent aerial agent for 3D bounding box annotation. The AGI-Coop dataset and its codebase will be made publicly available to facilitate future research in cooperative perception, autonomous driving and air-ground collaboration.

This report summarizes the system development, data collection and processing, and dataset generation efforts, which align with the objectives of year one. Based on the data collected and

processed in year-one, in year-two we will conduct the costs and benefit analysis of the cooperative system from three perspectives: safety, mobility, and environment.

2. System Development

2.1 Sensor Specifications

The dataset was collected using two instrumented vehicles, one roadside infrastructure, and two UAVs. Each vehicle carried a roof-mounted LiDAR, exterior cameras, and a GNSS/IMU system. The roadside infrastructure [7] integrated 32- and 64-beam LiDARs with complementary viewing angles, a video camera, and a long-range radar. Both UAVs used high-resolution cameras to record downward-facing 4K video footage. Detailed sensor models and specifications are provided in Table 1.

Table 1 Sensor configuration of the air–ground–infrastructure data-collection system.

Platform	Sensor	Model	Specifications and Configurations
Veh. I	LiDAR	Ouster OS1-64	64 beams; 360◦ horizontal and 45◦ vertical FoV; up to 120 m range; dual-return mode; 10 Hz
Veh. I	Camera	Lucid Triton RGB	Roof-mounted exterior RGB camera
Veh. I	Radar	Continental ARS 408-21	250 m maximum sensing range; horizontal FoV up to ±60◦
Veh. I	GNSS	OxTS OEM1000	RTK-enabled GNSS receiver for vehicle localization and timing
Veh. I	GPS/IMU	Xsens MTi	Integrated GPS/IMU for vehicle orientation and motion estimation
Veh. II	LiDAR	Velodyne VLP-32C	32 beams; 360◦ horizontal and 40◦ vertical FoV; up to 200 m range; dual-return mode; 10 Hz
Veh. II	Front camera	Mako G-319C	Front-facing RGB camera with a Sony IMX265 sensor
Veh. II	Rear camera	Logitech C920x	Rear-facing RGB camera
Veh. II	Radar	Aptiv ESR 2.5	174 m maximum sensing range; approximately ±10° FoV
Veh. II	GNSS	NovAtel Level 2.5 Kit	GNSS system for vehicle localization and time synchronization
Infra	LiDAR	Velodyne, 32-beam	32 beams; 360° horizontal FOV and −30° to +10° vertical FoV; 10 Hz
Infra	LiDAR	Ouster, 64-beam	64 beams; 360° horizontal FOV and −11.25° to 0° vertical FoV; 10 Hz
Infra	Camera	VIVOTEK FE9381-EHV	5 MP fisheye camera; 1920 × 1920 at 30 fps; 1.47 mm lens with 180° panoramic or 360° surround view
Infra	Radar	Econolite EVO	110° horizontal FOV and 60° maximum angle to traffic flow; 213-274 m maximum sensing range

Hovering UAV	Camera	DJI Matrice 4 enterprise	20 MP RGB camera with RTK positioning capability
Escort UAV	Camera	DJI Mavic 3 Classic	20 MP 4/3-inch CMOS camera

FoV: field of view. The vertical FoVs of the roadside LiDARs describe their configured viewing regions during data collection. Manufacturer specifications may differ from the actual acquisition settings.

2.2 Data Collection

As illustrated in Figure 1, we collected the data at a signalized urban intersection in Lafayette, Indiana. The roadside infrastructure was positioned at the northeast corner, while the two instrumented vehicles repeatedly traversed the intersection along predefined routes. By varying their maneuvers, such as left and right turns, straight movements, and stops at the traffic signal, we designed 30 distinct interactions, covering car-following, opposing approaches, simultaneous turns, crossing movements, and other coordinated vehicle encounters. The UAVs operated in two complementary modes: The escort UAV followed Vehicle I at an altitude of approximately 70 m above ground level to capture mobile aerial video, whereas the hovering UAV remained approximately stationary at 122 m (400 ft) above the intersection to provide persistent overhead coverage. We define the hovering UAV's coverage footprint, a rectangular area of approximately 175 m × 100 m, as the study area. The data collection was conducted continuously on a weekday afternoon and continued into the evening (3pm - 6pm), covering the complete duration of the evening peak traffic. After careful curation and synchronization, the dataset contains more than 6k multi-agent timestamps, comprising approximately 18k LiDAR frames and 36k camera images.

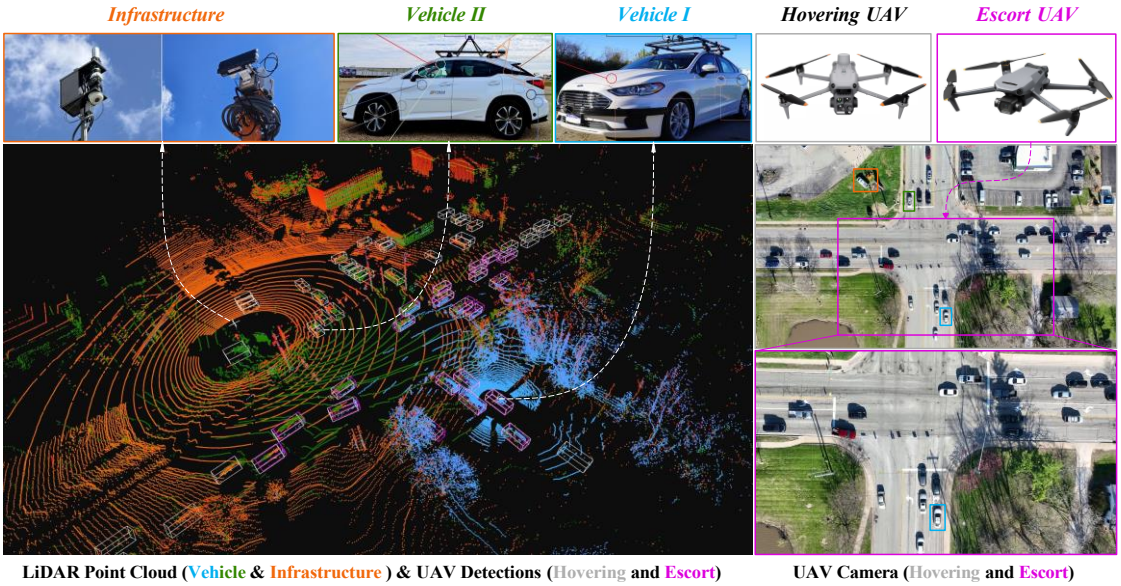


Figure 1 Demonstration of data acquisition systems.

The orange, green, and blue boxes in the UAV camera images represent the location of the roadside infrastructure, instrumented vehicle II, and instrumented vehicle I, respectively.

2.3 UAV Detection and Tracking

Given the limited prior work on intermediate fusion between UAV imagery and ground-sensor observations, we demonstrate the value of UAV detections through late fusion with detections produced by the ground sensors. For the hovering UAV, vehicle detections are extracted from the UAV footage using the open-source Geo-Trax framework [8], which couples a pretrained YOLO detector with BoT-SORT multi-object tracking [9]. Each video frame is registered to a master reference frame to mitigate the effects of UAV hovering instability. The resulting vehicle detections are shown in the left panel of Figure 2.

For the escort UAV, continuous changes in camera position and orientation produce a time-varying image-to-map transformation throughout the flight. Although YOLO and BoT-SORT can still generate image-space vehicle tracks, converting them into map-referenced trajectories requires a reliable transformation. A single master reference is therefore insufficient, because simply chaining frame-to-frame transformations would accumulate drift. We address this challenge with a telemetry-assisted, anchor-based method that constructs a frame-wise image-to-orthophoto transformation sequence and projects the tracked vehicles into the Universal Transverse Mercator (UTM) coordinate system, Zone 16N, based on the WGS 84 datum.

Frame-level telemetry embedded in the video subtitle track is first aligned in time with the decoded onboard flight log. Sparse anchor frames are then registered directly to the site orthophoto at an initial interval of approximately 5 s and adaptively densified across long registration gaps and substantial platform rotations. Vehicle detections are masked during feature matching, while the GNSS-derived UAV position and short-horizon gimbal-heading changes provide spatial and rotational constraints on candidate registrations. Candidates are retained only if they satisfy geometric and matching quality gates on anisotropy, perspective distortion, projected-footprint shape, feature-inlier support, and reprojection error.

The accepted anchors are used to construct the framewise transformation sequence. For frames between anchors, rotation is interpolated from the logged heading changes, while translation follows the smoothed UAV trajectory. When neighboring projective anchors are not mutually consistent, the sequence is rebuilt using a similarity model with rotation, uniform scale, and two-dimensional translation. We fit this similarity to the full homography's action over a uniform image grid [10], rather than extracting rotation from its anisotropic linear component, which can introduce systematic bias.

The resulting transformation sequence is evaluated using inter-frame visual-odometry consistency and an independent comparison of road-axis orientation against the orthophoto. Each tracked detection is then projected through its frame-specific transformation to obtain

UTM coordinates for subsequent trajectory refinement. The detections of the escort UAV are shown in the right panel of Figure 2.

The raw tracks from both the hovering and escort UAV detections are further refined using a post-processing pipeline that denoises vehicle positions and dimensions, interpolates short detection gaps, and suppresses false positives. For moving vehicles, yaw is re-estimated from the motion trajectory because Geo-trax provides only axis-aligned yaws. The resulting per-frame vehicle state consists of a 2-D position $c_{\text{uav}} = (x_{\text{uav}}, y_{\text{uav}})$, planar dimensions $d_{\text{uav}} = (L_{\text{uav}}, W_{\text{uav}})$, a yaw angle, and a semantic class from $(Car, Truck, Bus, Motorcycle)$.



Hovering UAV **Escort UAV**

Figure 2 Demonstration of UAV detections.

2.4 Coordinate System

The sensor measurements and detections from all agents are expressed in a shared map frame defined by a preconstructed FAST-LIO2 LiDAR map [11], as shown in Figure 3. The LiDAR scans from the three ground agents are registered to this map using GNSS-initialized iterative closest point (ICP) [12], yielding the rigid transformation from each LiDAR coordinate frame to the shared map frame. In addition, the camera-to-LiDAR extrinsic parameters of each ground agent are calibrated offline, allowing camera observations to be transformed into the same map frame.

The UAV detections are aligned with the shared map frame through two mapping products: an orthophoto and a photogrammetric point cloud (Figure 3). Both are generated from overlapping UAV images of the study area using a structure-from-motion (SfM)-based photogrammetry pipeline. A digital terrain model (DTM) is further derived from the point cloud to provide terrain elevation. Each UAV detection is first transformed from image coordinates to UTM coordinates using the orthophoto. Its ground elevation is then obtained from the DTM at the corresponding planar UTM location. The photogrammetric point cloud is registered to the LiDAR map using ICP, yielding a rigid transformation from the UTM frame to the shared map frame. By composing these transformations, the UAV detections are expressed in the shared map frame and can subsequently be transformed into the sensor frames of the ground agents.

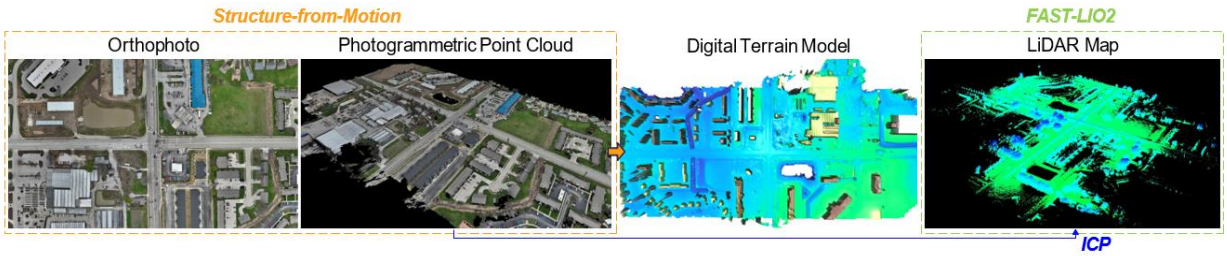


Figure 3 Demonstration of mapping products

2.5 Data Annotation

Data annotations are initialized from the vehicle detections of the hovering drone and then refined using the combined point clouds of the two vehicles and the infrastructure. For each vehicle track, dimensions are first estimated from the temporally aggregated LiDAR points cropped around the corresponding detections to have a complete picture. The bounding box pose and dimensions are then refined frame by frame whenever sufficient LiDAR evidence is available. Each annotation provides the vehicle's 3D position $c_{GT} = (x_{GT}, y_{GT}, z_{GT})$, dimensions (L_{GT}, W_{GT}, H_{GT}), yaw angle, and class $(Car, Truck, Bus, Motorcycle)$ inherited from Hovering UAV detections. Our dataset has over 400K 3D bounding boxes, on average 65.3 objects per frame with a maximum of 86 and a minimum of 37, which represents a dense traffic scenario.

Besides manual annotation, we also develop DAA (Drones as Annotators), a drone-assisted framework for unsupervised 3D auto-labeling. DAA augments onboard LiDAR with the hovering UAV. From its bird's-eye view (BEV), the UAV detects and tracks vehicles over a broad area with substantially less occlusion, which complements cases exactly where the onboard LiDAR fails. In turn, the LiDAR unit supplies precise metric 3D geometry and vertical structure that could not be recovered reliably from aerial imagery alone. Through an iterative refinement process, DAA exploits the two complementary measurements that alternates between identifying the foreground LiDAR points associated with each object under the current bounding box estimation and updating the box using temporally aggregated LiDAR points together with the aerial prior. More details on DAA can be found in [13].

3. Findings

We evaluate the performance of different cooperative perception settings through combinations of sensor modalities in two categories: (1) ground-level cooperative perception and (2) air-ground-infrastructure cooperative perception, which are detailed in Table 2.

Table 2 Cooperative perception settings evaluated in AGI-Coop. Veh. I is the ego vehicle in all multi-agent configurations.

Setting	Description
Ground-level cooperative perception	
Veh. I	Ego-only perception using observations from Instrumented Vehicle I.
Veh. I + Veh. II	V2V cooperation, where Instrumented Vehicle II provides a complementary mobile viewpoint to Vehicle I.
Veh. I + Infra.	V2I cooperation, where roadside sensing augments the ego vehicle with a fixed viewpoint.
Veh. I + Veh. II + Infra.	V2X cooperation combining observations from both instrumented vehicles and the roadside infrastructure.
Air-Ground-Infrastructure Cooperative Perception	
Escort UAV	Standalone perception from the escort UAV, which follows Vehicle I and provides mobile aerial coverage.
Hovering UAV	Standalone perception from the hovering UAV, which provides persistent overhead coverage of the study area.
Veh. I + Escort UAV	V2A cooperation combining Instrumented Vehicle I's perception with mobile aerial observations from the escort UAV.
Veh. I + Veh. II + Escort UAV	V2V2A cooperation combining two vehicle viewpoints with observations from the escort UAV.
Veh. I + Infra. + Escort UAV	V2I2A cooperation combining the ego vehicle, roadside infrastructure, and the escort UAV.
Veh. I + Veh. II + Infra. + Escort UAV	Full cooperation integrating two vehicles, roadside infrastructure, and the escort UAV.

3.1 Implementation Details

We divide the dataset into training and validation sets using the 80:20 split. Detection performance is evaluated within a rectangular region centered on the ego vehicle, covering ([-110.4, 110.4]) m along both the (x)- and (y)-axes. All models are trained for 40 epochs on four NVIDIA A6000 GPUs with a batch size of 2. We employ the Adam optimizer with an initial learning rate of 0.002. The learning rate is decayed by a factor of 0.1 at the 15th and 25th epochs. We benchmark four representative fusion paradigms: **No Fusion**, the instrumented vehicle I independently performs 3D object detection using PointPillars, without receiving information from other agents. **Early Fusion**, the raw point clouds collected by the participating agents are first transformed into the ego-vehicle coordinate system and subsequently concatenated. The aggregated point cloud is then processed by a unified PointPillars network for end-to-end training and inference. **Intermediate Fusion**, each agent independently extracts intermediate features using its corresponding feature encoder. The resulting features are spatially aligned in the ego vehicle coordinate system and subsequently

aggregated using different feature-level fusion strategies. All intermediate fusion experiments are implemented using the OpenCOOD framework. **Late Fusion**, each agent independently generates 3D bounding-box predictions using PointPillars.

The predictions are then transformed into the ego-vehicle coordinate system and merged during the non-maximum suppression (NMS) stage. To account for heterogeneous sensing characteristics of agents from different domains, sensor- and domain-specific confidence weights are introduced to select and refine overlapping bounding boxes.

3.2 Quantitative Results

Ground-level Cooperative Perception

Table 3 compares representative fusion strategies under the ground-level V2V and V2I cooperative perception. Late fusion consistently achieves the best performance across all settings, reaching 87.19%, 82.98%, and 47.05% at AP@0.3, AP@0.5, and AP@0.7, respectively, under the full V2X configuration. V2I cooperation also provides substantially larger gains than V2V cooperation. For late fusion, the infrastructure agent improves AP@0.5 from 61.10% to 75.88%, demonstrating the strong complementarity of its elevated and spatially distinct viewpoint. Among the intermediate fusion approaches, HEAL and CoAlign exhibit the most competitive performance, whereas several methods degrade when heterogeneous agents are jointly introduced. This result indicates that increasing the number of collaborators does not necessarily guarantee better performance and highlights the challenges of cross-agent feature alignment and heterogeneous representation fusion.

Table 3 Comparison of Fusion Strategies under Canonical V2V, V2I, and V2X Topologies on the AGI-COOP Dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Methods	V2V (Veh. I + Veh. II)			V2I (Veh. I + Infra.)			V2X (Veh. I + Veh. II + Infra.)		
	AP@0.3	AP@0.5	AP@0.7	AP@0.3	AP@0.5	AP@0.7	AP@0.3	AP@0.5	AP@0.7
No Fusion	38.96	33.40	13.95	38.96	33.40	13.95	38.96	33.40	13.95
Early Fusion	54.24	47.56	19.27	65.52	59.05	24.65	76.45	71.19	33.55
Late Fusion	**68.26**	**61.10**	**27.82**	**81.18**	**75.88**	**39.73**	**87.19**	**82.98**	**47.05**
AttFusion	40.50	29.24	8.85	59.25	39.25	9.31	55.07	34.65	7.06
F-Cooper	46.78	32.03	9.43	54.47	35.97	7.12	49.28	28.80	3.85
CoBEVT	49.98	34.59	11.74	60.69	39.11	11.67	57.78	35.35	9.27
V2XViT	50.50	35.87	12.01	67.76	51.48	15.86	75.18	61.34	22.44
CoAlign	58.28	47.00	15.14	74.49	66.30	33.44	79.94	73.04	35.92
HEAL	59.67	49.29	18.04	79.00	72.33	39.61	82.82	75.09	38.67

Table 4 Scalability of Ego-Centric Cooperative Perception under Different Vehicle-Infrastructure-Aerial Agent Configurations. The performance gains are calculated relative to the Ego-only baseline.

Cooperative Agents	Absolute Performance			Performance Gain (Δ vs. Ego-Only)		
	AP@0.3	AP@0.5	AP@0.7	ΔAP@0.3	ΔAP@0.5	ΔAP@0.7
Veh. I (Ego-Only)	38.96	33.40	13.95	-	-	-
Escort UAV (Only)	26.29	20.87	2.21	-	-	-
Hovering UAV (Only)	74.18	66.29	18.70	-	-	-
Veh. I + Escort UAV	45.96	37.89	15.72	+7.00	+4.49	+1.77
Veh. I + Veh. II + Escort UAV	71.53	63.36	29.14	+32.57	+29.96	+15.19
Veh. I + Infra. + Escort UAV	82.19	76.89	40.28	+43.23	+43.49	+26.33
Veh. I + Veh. II + Infra. + Escort UAV	**87.73**	**83.56**	**47.51**	**+48.77**	**+50.16**	**+33.56**

Air-Ground-Infrastructure Cooperative Perception

Table 4 further investigates the contribution and scalability of aerial cooperation under the late fusion setting. The escort UAV alone achieves limited performance because of its sparse and dynamically moving observations, whereas the hovering UAV obtains an AP@0.5 of 66.29%, substantially outperforming the ego-only baseline of 33.40%. When combined with the ego vehicle, the escort UAV provides a consistent AP@0.5 gain of 4.49 percentage points, and its benefit becomes more pronounced when additional vehicle and infrastructure agents are incorporated. The complete four-agent configuration achieves the best performance, with AP@0.3, AP@0.5, and AP@0.7 scores of 87.73%, 83.56%, and 47.51%, corresponding to absolute improvements of 48.77, 50.16, and 33.56 percentage points over the ego-only baseline. These results demonstrate that vehicle-, infrastructure-, and aerial-views provide complementary spatial coverage, while also revealing the considerable performance difference between moving escort and stationary hovering UAV configurations.

3.3 Qualitative Results

As shown in Figure 4, the three fusion paradigms exhibit clear differences in detection completeness and spatial coverage. Early Fusion detects most objects near the ego vehicle but misses several targets in distant or weakly observed regions. HEAL improves the coverage by aggregating intermediate features, yet some objects remain undetected or inaccurately localized under the heterogeneous V2I and V2X configurations. In comparison, Late Fusion produces more complete detections and better preserves objects observed exclusively by collaborating agents. The improvement is particularly evident in the peripheral and occluded regions, demonstrating the complementary benefits of multi-view observations.

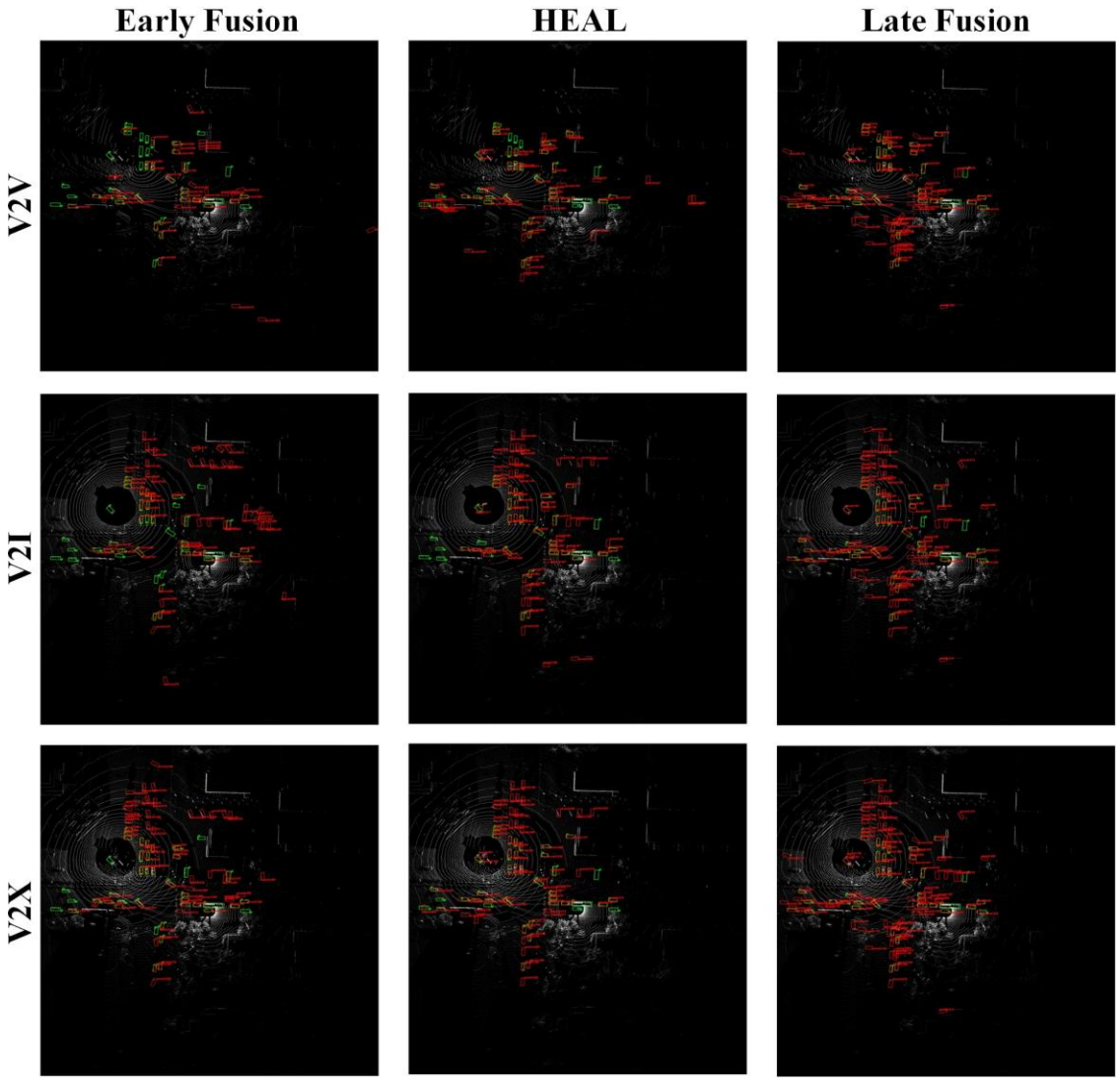


Figure 4 Qualitative comparison of Early Fusion, HEAL, and Late Fusion under V2V, V2I, and V2X cooperative topologies.

The columns correspond to different fusion methods, while the rows represent different cooperative settings. The green and red bounding boxes denote ground-truth annotations and detection results, respectively.

4. Recommendations

The AGI-coop dataset has great potential to test, validate and evaluate new methods across multiple research domains. Below we summarize recommendations on how this dataset can be utilized beyond cost-benefit analysis.

Multimodal Cooperative Perception. The combination of aerial, roadside, and vehicle sensors is

particularly valuable for investigating cross-view perception models since it contains both ground-level and aerial views of the same scene. It can therefore be used to quantify how viewing angle, sensing range, object distance, and occlusion affect the perception performance. The heterogeneous sensor suite also enables research on multimodal fusion at different levels.

Cooperative Perception-based Autonomous Driving and ITS Applications.
For autonomous driving research, this dataset provides a test environment for evaluating the value of different cooperative perception settings and sensor fusion techniques. The performance of the ego vehicle's downstream tasks (e.g., trajectory planning) using only its onboard sensors can be compared with raw data or processed detections from another vehicle, roadside infrastructure, and one or both UAVs. Similarly, different settings and fusion techniques can be integrated with various cooperative perception-based ITS applications such as traffic signal optimization, occlusion-aware collision warning, and traffic state estimation.

Ground-Air Cooperation. The two UAVs with different operating modes create opportunities to study ground-air mobility coordination. The hovering UAV provides a stable and global view, which could be utilized to optimize traffic flow and vehicle trajectories of the entire scene, such as UAV-assisted cooperative driving and autonomous intersection management (AIM) with the UAV as the coordinator/scheduler. On the other hand, the escort UAV provides a moving viewpoint that remains associated with a particular ground vehicle. This following behavior could extend the perception and detection range of the ground vehicle and enable various collaboration mechanisms. For example, if the ground vehicle is an autonomous vehicle, the additional aerial observations can reveal objects and traffic conditions beyond the range or line of sight of its onboard sensors, thereby improving situational awareness, trajectory prediction, and motion planning. If the ground vehicle is a human-driven vehicle, the UAV can provide additional information such as advance warnings of hazards, downstream queue locations, or vehicles approaching from occluded areas.

5. Outputs, Outcomes, and Impacts

The major outcome of this project is the development of the AGI-Coop real-world air–ground–infrastructure cooperative perception dataset and two research papers submitted to the 2027 Transportation Research Board Annual Meeting. The dataset and associated codebase are intended for public release, allowing researchers and transportation agencies to evaluate multimodal perception methods and their application to autonomous driving and ITS applications. The results demonstrate the practical value of combining vehicle, roadside, and aerial observations. Although the project did not directly implement changes to transportation operations, regulations, legislation, or policy during the reporting period, its outputs provide an evidence base and reusable research infrastructure for developing more reliable and cost-

effective cooperative perception systems and evaluating research outcomes related to cooperative perception.

The following outputs were generated during the performance of this project

- Two papers submitted for presentation at 2027 TRB Annual Meeting
  Zhu, T., Wang, Z., and Y. Feng, Drones as Annotators: Amodal 3D Auto-Labeling for Ground LiDAR with Aerial Priors. Submitted to 2027 Transportation Research Board Annual Meeting.
  Zhu, T. et al. AGI-Coop: A Real World Air-Ground-Infrastructure Cooperative Perception Dataset. Submitted to 2027 Transportation Research Board Annual Meeting.

References

[1] Cao, P., Xiong, Z., Liu, X., 2022. An analytical model for quantifying the efficiency of traffic data collection using instrumented vehicles. Transportation research part C: emerging technologies 136, 103558.
[2] Chen, H., Liu, B., Zhang, X., Qian, F., Mao, Z.M., Feng, Y., 2022. A cooperative perception environment for traffic operations and control. arXiv preprint arXiv:2208.02792 .
[3] Xu, R., Xiang, H., Han, X., Xia, X., Meng, Z., Chen, C. J., ... & Ma, J. (2023). The OpenCDA opensource ecosystem for cooperative driving automation research. IEEE Transactions on Intelligent Vehicles.
[4] Zhang, R., Meng, D., Shen, S., Zou, Z., Li, H., & Liu, H. X. (2023). MSight: An Edge-Cloud Infrastructure-based Perception System for Connected Automated Vehicles. arXiv preprint arXiv:2310.05290.
[5] Xu, R., Xia, X., Li, J., Li, H., Zhang, S., Tu, Z., ... & Ma, J. (2023). V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13712-13722).
[6] Xiang, H., Zheng, Z., Xia, X., Xu, R., Gao, L., Zhou, Z., ... & Ma, J. (2024, September). V2x-real: a large-scale dataset for vehicle-to-everything cooperative perception. In European Conference on Computer Vision (pp. 455-470). Cham: Springer Nature Switzerland.
[7] Tarko, A. P., K. B. Ariyur, M. A. Romero, V. K. Bandaru, and C. G. Lizarazo, Tscan: Stationary Lidar for traffic and safety studies—object detection and tracking, 2016.
[8] Fonod, R., H. Cho, H. Yeo, and N. Geroliminis, Advanced computer vision for extracting georeferenced vehicle trajectories from drone imagery. Transportation Research Part C: Emerging Technologies, Vol. 178, 2025, p. 105205.
[9] Aharon, N., R. Orfaig, and B.-Z. Bobrovsky, BoT-SORT: Robust associations multi-pedestrian tracking. arXiv:2206.14651, 2022.
[10] Umeyama, S., Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 4, 1991, pp. 376–380.
[11] Xu, W., Y. Cai, D. He, J. Lin, and F. Zhang, FAST-LIO2: Fast Direct LiDAR-Inertial Odometry. IEEE Transactions on Robotics, Vol. 38, No. 4, 2022c, pp. 2053–2073.
[12] Besl, P. J. and N. D. McKay, Method for registration of 3-D shapes. In Sensor fusion IV: control paradigms and data structures, SPIE, 1992, Vol. 1611, pp. 586–606.
[13] Zhu, T., Wang, Z., and Y. Feng, Drones as Annotators: Amodal 3D Auto-Labeling for Ground LiDAR with Aerial Priors. Submitted to 2027 Transportation Research Board Annual Meeting.