

# CoPeD-Advancing Multi-Robot Collaborative Perception: A Comprehensive Dataset in Real-World Environments

Yang Zhou<sup>1</sup>, Graduate Student Member, IEEE, Long Quang<sup>2</sup>, Carlos Nieto-Granda<sup>1</sup>,  
and Giuseppe Loianno<sup>1</sup>, Member, IEEE

**Abstract**—In the past decade, although single-robot perception has made significant advancements, the exploration of multi-robot collaborative perception remains largely unexplored. This involves fusing compressed, intermittent, limited, heterogeneous, and asynchronous environmental information across multiple robots to enhance overall perception, despite challenges like sensor noise, occlusions, and sensor failures. One major hurdle has been the lack of real-world datasets. This letter presents a pioneering and comprehensive real-world multi-robot collaborative perception dataset to boost research in this area. Our dataset leverages the untapped potential of air-ground robot collaboration featuring distinct spatial viewpoints, complementary robot mobilities, coverage ranges, and sensor modalities. It features raw sensor inputs, pose estimation, and optional high-level perception annotation, thus accommodating diverse research interests. Compared to existing datasets predominantly designed for Simultaneous Localization and Mapping (SLAM), our setup ensures a diverse range and adequate overlap of sensor views to facilitate the study of multi-robot collaborative perception algorithms. We demonstrate the value of this dataset qualitatively through multiple collaborative perception tasks. We believe this work will unlock the potential research of high-level scene understanding through multi-modal collaborative perception in multi-robot settings.

**Index Terms**—Data sets for robotic vision, deep learning for visual perception, multi-robot systems.

## I. INTRODUCTION

ROBOTIC perception has advanced significantly over the past decade, largely due to progress in machine learning and the widespread adoption of increasingly powerful computing and sensor systems. However, most of these developments

have been centered on single-robot systems, leaving the area of multi-robot collaborative perception relatively unexplored. This form of perception refers to a network of robots, each with its unique sensing characteristics, capabilities, and spatial viewpoints being able to collect and fuse environmental sensing data in an efficient, distributed, and meaningful way to obtain accurate context-appropriate information, leading to potential advantages in terms robustness, accuracy, and redundancy compared to a single robot solution. Pushing the frontier of multi-robot collaborative perception could harness these unique advantages to create a more comprehensive, unified, and resilient perception system. The creation of multi-robot perception datasets can support and facilitate the development of collaborative perception models and algorithms. In this paper, we present the first heterogeneous multi-robot perception dataset collected in challenging indoor and outdoor settings using ground and aerial robots.

Notably, the multi-robot air-ground systems have the potential to provide a distinct opportunity for multi-robot perception. Specifically, ground and small-scale aerial robots combine different mobility and sensing characteristics (e.g., types of sensors as well as different spatial viewpoints and sensing range characteristics). For instance, ground robots can carry heavy payloads such as 3D LiDAR, whereas small-scale aerial robots due to their Size, Weight, and Power (SWaP) constraints generally can only rely on small sensors such as cameras and Inertial Measurements Units (IMUs).

Existing multi-robot datasets are mainly geared towards Simultaneous Localization and Mapping (SLAM) applications [1], [2], [3], [4], [5], focusing primarily on spatial coverage rather than collaborative perception. The limited overlap of sensor views in these datasets hinders the study of perception results that rely on perception compensation mechanisms between robots. Furthermore, most of the existing multi-agent perception datasets are based on photorealistic simulators, neglecting the real-world sensor noise and characteristics inherent in real-world multi-robot settings. Although there are datasets targeting autonomous driving systems through vehicle-to-vehicle (V2V) communication [6], [7], [8], [9] these fall short due to the mismatch between simulated and real-world data or the absence of multiple vehicles with heterogeneous mobility and modalities. Furthermore, the algorithms developed from these datasets are designed for Bird's Eye View (BEV) since the

Manuscript received 7 March 2024; accepted 21 May 2024. Date of publication 28 May 2024; date of current version 5 June 2024. This letter was recommended for publication by Associate Editor Gennaro Notomista and Editor M. Ani Hsieh upon evaluation of the reviewers' comments. This work was supported in part by DARPA YFA under Grant D22AP00156-00 and Grant DEVCOM ARL SARA W911NF-24-2-0057 and in part by NSF CPS under Grant CNS-2121391. (Corresponding author: Yang Zhou.)

Yang Zhou and Giuseppe Loianno are with the New York University, Tandon School of Engineering, Brooklyn, NY 11201 USA (e-mail: yangzhou@nyu.edu; loiannog@nyu.edu).

Long Quang and Carlos Nieto-Granda are with the U.S. Army Combat Capabilities Development Command, Army Research Laboratory, Adelphi, MD 20783 USA (e-mail: long.p.quang.civ@army.mil; carlos.p.nieto2.civ@army.mil).

This letter has supplementary downloadable material available at <https://doi.org/10.1109/LRA.2024.3406207>, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3406207

TABLE I  
TAXONOMY OF EXISTING DATASETS

Task	Approaches	Platform Type	Real-World	Modalities	Scene
Multi-Robot SLAM	Lamp 2.0 [1]	Multi-Ground Robot	✓	Camera, LiDAR	Outdoor
	Subt-Cerberus [2]	Multi-Ground Robot	✓	Camera, LiDAR, IMU	Outdoor
	Graco [3]	Aerial/Ground Robot	✓	Camera, LiDAR, IMU, GPS	Outdoor
	S3e [4]	Aerial/Ground Robot	✓	Camera, LiDAR, IMU, GPS	Indoor/Outdoor
	Tian et al. [5]	Multi-Ground Robot	✓	Camera, LiDAR, IMU, GPS	Outdoor
Collaborative Perception	Opv2v [6]	Vehicle-to-Vehicle (V2V)	✗	Camera, LiDAR, IMU, GPS	Outdoor
	Arnold et al. [7]	Vehicle-to-Vehicle (V2V)	✗	Camera, Depth Camera	Outdoor
	Dair-V2X [8]	Vehicle-to-Everything (V2X)	✓	Camera, LiDAR, IMU, GPS	Outdoor
	V2X-Sim [9]	Vehicle-to-Vehicle (V2V)	✗	Camera, LiDAR, IMU, GPS	Outdoor
	When2comm [10]	Multi-Aerial Robot	✗	Camera	Outdoor
	Where2comm [11]	Multi-Aerial Robot	✗	Camera	Outdoor
	CoPeD (Ours)	Multi-Aerial/Ground Robot	✓	Camera, LiDAR, IMU, GPS	Indoor/Outdoor

scene is constrained on the flat surface and agent mobility is limited to ground vehicle. Due to the limited scope and biases of existing datasets, there is no existing methods that utilize multiple modalities, multiple mobilities and goes beyond BEV paradigm. Creating a comprehensive dataset tailored to address multi-robot collaborative perception problems, a crucial yet unexplored area in the community, is indispensable to making significant progress in this area and push the frontier to develop new methods.

The contributions of this paper can be summarized as follows. First, we present the first multi-robot dataset tailored to collaborative perception tasks featuring both indoor and outdoor sequences, multiple sensor spatial viewpoints, robot types, multi-rate and multi-modal data as well as coverage ranges. It reflects the real-world sensor noise and disturbances encountered when deploying a heterogeneous robot team in the wild. The aerial robots equipped with stereo cameras, forward and downward RGB cameras, depth sensors, Inertial Measurement Units (IMUs), and GPS. Ground robots carry a stereo camera, forward RGB camera, depth sensor, 3D LiDAR, IMU, and GPS. Second, to cater to diverse research needs, in addition to raw sensor data streams, we provide pose estimation and optional high-level perception annotation. Specifically, pose estimation is obtained by fusing GPS data with existing SLAM frameworks. For high-level perception annotation, we employ foundation models to obtain zero-shot automatic annotation. Finally, we validate the usefulness of the dataset in several collaborative perception tasks. Our dataset is a first-of-its-kind contribution to the robotics community going beyond the BEV autonomous driving scenario. The incorporation of multiple sensor modalities and robots enables to pave the way for studying multi-modal collaborative perception using real-world data.

The paper is organized as follows. In Section II, we discuss existing datasets. In Section III, we present the dataset characteristics, and in Section IV the technical details related to sensors and calibration. Section V introduces the pose and perception annotations procedures, Section VI discuss the dataset attributes, whereas Section VII presents use cases and applications. Finally, Section VIII concludes the work and presents several future research directions.

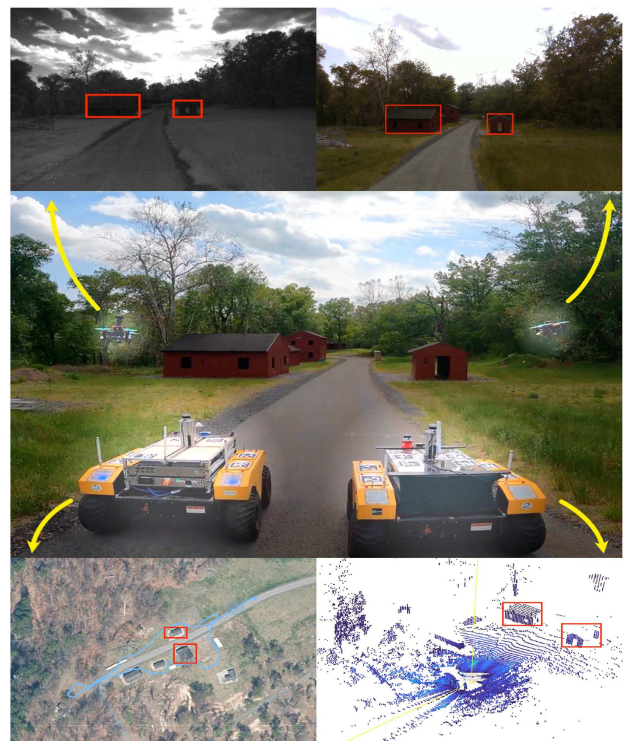


Fig. 1. Multi-robot collaborative data collection. The top row shows the infrared and RGB images from the aerial robots, whereas the bottom row shows the GPS and LiDAR data from the ground robots. The sensor data present several overlapping spatial areas. Red boxes show sample objects identified in multiple heterogeneous sensors' data streams.

## II. RELATED WORKS

This section reviews existing datasets and highlights the differences between our approach and existing solutions as illustrated in Table I. The field of multi-robot collaborative perception is relatively new [12], [13], with few datasets available for research purposes within the community. Existing datasets are primarily designed for SLAM applications [1], [2], [3], [4], [5], while our dataset is specifically tailored to facilitate research into multi-robot collaborative perception.

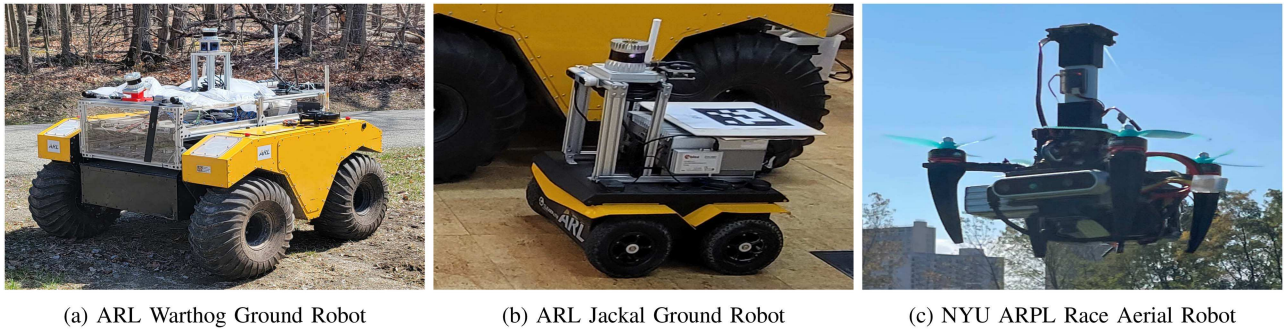


Fig. 2. Layout of the ground robots and aerial robots.

Historically, the multi-robot collaborative perception was primarily confined to geometric mapping in robotics. However, significant advancements have been made in this area by the research community [14], resulting in the release of several multi-robot SLAM datasets [1], [2], [3], [4], [5]. These datasets contain outdoor sequences, with the exception of [4], which includes both indoor and outdoor sequences. Datasets such as [1], [4], [5] deploy multiple ground robots for data collection, and [2], [3] extend this to air-ground robot teams, but only in outdoor urban settings. Despite these advancements, these datasets are primarily designed for SLAM applications, focusing on geometric mapping rather than high-level collaborative perception tasks, including monocular depth estimation [12], semantic segmentation [10], [12], [15], [16], correspondence matching [17], object detection [11], [13], [18], [19], [20], [21]. The limited overlap of sensor views and types of environments in these datasets hinders the study of perception models and algorithms that exploit the benefits of using multiple robots.

There are a few collaborative perception datasets target autonomous driving systems through V2V communication [6], [7], [8], [9]. Specifically, most of these datasets, due to the difficulties associated with the collection and annotation of real-world data, are based on simulation with the exception of [8] which provides the first real-world V2X dataset. The dataset from [7] employs LiDAR as the single sensing modality, while [6], [8], [9] adopt multiple modalities including cameras and 3D LiDARs. However, these datasets are intended for autonomous driving applications, missing the aerial complementary viewpoints and perspectives, and focusing on 3D object detection tasks. In addition, most of these datasets are based on simulation due to the difficulties associated with real-world data collection and annotation. In addition to autonomous driving-focused datasets, [10] and [11] provide simulation-based photorealistic datasets from multiple aerial robot perspectives. Therefore, it is important to acknowledge that this data alone cannot completely replicate the nuanced aspects of robot data available in the real-world. Finally, existing datasets do not fully offer the benefits of data streams from multiple distinct spatial complementary viewpoints and perspectives as in the proposed air-ground case.

Datasets need to be more comprehensive to be used for collaborative perception tasks. Very few are designed for heterogeneous multi-robot applications, and the sensor modality

setups are unsuitable for multi-robot collaborative perception. Compared to existing datasets as summarized in Table I, the proposed dataset is the first real-world heterogeneous air-ground collaborative perception dataset, designed explicitly for multi-modal collaborative robot perception research.

### III. DATASET DESCRIPTION

#### A. Robots and Equipment

Our dataset utilizes three ground robots and two aerial robots. The robot layouts are illustrated in Fig. 2. The Clearpath Robotics Warthog platform is equipped with one Ouster OS1-64 3D LiDAR, a LORD Microstrain 3DM-GX5-25 IMU, an u-Blox EVK-M8T GNSS module, an Intel Realsense D435i/D455 stereo camera, two high-resolution FLIR Blackfly S cameras and a hardware time synchronization Masterclock GMR1000. The Clearpath Robotics Jackal platform is equipped with an Ouster OS1-64 3D LiDAR, a LORD Microstrain 3DM-GX5-25 IMU, an u-Blox EVK-M8T GNSS module, and an Intel Realsense D435i/D455 stereo camera. Each of the aerial robots, developed at the Agile Robotics and Perception Lab (ARPL)<sup>1</sup> at the New York University, carries a forward Intel Realsense D435i stereo camera, a downward IMX219 color camera, a top-mounted PX4 autopilot module and an mRO Location One GPS module. The D435i stereo camera includes an RGB camera module, a stereo infrared camera module, and an IMU module. The complete list of sensor specifications is shown in Tables II and III. All robots' autonomy software stack builds upon ROS 1 [22]<sup>2</sup>, and all robots are connected under one 5 G Wi-Fi subnetwork with time synchronization to a laptop's system clock using Network Time Protocol (NTP).

We mount AprilTag [23] on ground robots illustrated in Fig. 2 to provide relative localization visual markers between the aerial robots and the ground robots. Compared to other visual fiducial markers, AprilTags guarantee increased robustness and accuracy from the perspectives of long-range detection and localization accuracy under real-time computation constraints. The aerial robots obtain the relative pose of the ground robots by AprilTag detection and Perspective-n-Point (PnP) [24] using downward cameras. We use 8 AprilTags as a tag bundle in order to facilitate

<sup>1</sup>[Online]. Available: <https://wp.nyu.edu/arp/>

<sup>2</sup>[Online]. Available: <http://wiki.ros.org/>

TABLE II  
GROUND ROBOT SENSOR SPECIFICATIONS

Equipment	Model Type	Characteristics	Resolution	FoV	Sensor Rate
LiDAR	Ouster OS1-64	200 m range	512 × 64	45° vertical	10 Hz
LiDAR	Ouster OS1-128	200 m range	512 × 128	45° vertical	10 Hz
LiDAR	Ouster OS0-64	100 m range	512 × 64	90° vertical	10 Hz
IMU	LORD Microstrain 3DM-GX5-25	±8 g	300 dps	-	1000 Hz
GNSS	u-Blox EVK-M8T	2 m, 0.3° accuracy	-	-	2 Hz
RGBD (Wilbur)	Intel Realsense D435i	RGB: Rolling Shutter	1920 × 1080	69° × 42°	30 Hz
		Stereo IR: Global Shutter, Baseline: 50 mm IMU: BMI 055	640 × 480	87° × 58°	30 Hz 200 Hz
RGBD (Wanda)	Intel Realsense D455	RGB: Global Shutter	1280 × 800	90° × 65°	30 Hz
		Stereo IR: RS, Baseline: 95 mm IMU: BMI 055	1280 × 720	87° × 58°	30 Hz 200 Hz
Stereo Camera	FLIR Blackfly S (BFS-PGE-16S2C-CS)	RGB: Global Shutter, Baseline: 60 cm	1440 × 1080	-	78 Hz
Sync	Masterclock GMR1000	±3 second / year	-	-	-
Main Computer	Nuvo-7166GC	Intel i7-8700 CPU @ 3.20GHz w. 32GB RAM, NVIDIA T4 GPU w. 16GB RAM	-	-	-
Vision Computer	Nuvo-7166GC	Intel i7-8700 CPU @ 3.20GHz w. 32GB RAM, NVIDIA T4 GPU w. 16GB RAM	-	-	-

TABLE III  
AERIAL ROBOT SENSOR SPECIFICATIONS

Equipment	Model Type	Characteristics	Resolution	FoV	Sensor Rate
GNSS	mRO Location One	ublox NEO-M9N	-	-	8 Hz
		Magnetic Compass RM3100	-	-	8 Hz
		Barometer DPS310	-	-	8 Hz
RGB	IMX219	RGB: Global Shutter	640 × 480	160° × 160°	30 Hz
RGBD	Intel Realsense D435i	RGB: Rolling Shutter	1920 × 1080	69° × 42°	30 Hz
		Stereo IR: Global Shutter, Baseline: 50 mm IMU: BMI 055	640 × 480	87° × 58°	30 Hz 200 Hz
Computer	NVIDIA Jetson Xavier NX	6-core NVIDIA Carmel ARM v8.2 64-bit CPU, 384-core NVIDIA Volta GPU with 48 Tensor Cores, 8 GB RAM	-	-	-

robust AprilTag detection in the presence of occlusion, illumination changes, motion blur, and FoV constraints. The tag bundle is calibrated and detected using the method in [25].

Our heterogeneous robot team setup ensures that a diverse array of sensor modalities and robot mobilities are captured, thus enriching the dataset and providing more robust material for multi-modal heterogeneous robotic perception studies.

### B. Indoor Sequences

For indoor sequences, we use one Jackal ground robot, one Warthog ground robot, and two aerial robots. We present three sequences with one Warthog ground robot and one aerial robot, and one sequence with two ground robots and two aerial robots as shown in Fig. 3.

The indoor sequences named 'Indoor-NYUARPL' are captured in a large indoor environment of 38~m × 60~m with a rich composition of objects including tables, chairs, and doors. We split the robot team into two subteams, each of them consisting of one aerial robot and one ground robot. The two subteams explore different parts of the environment. The aerial robots fly



Fig. 3. Indoor-NYUARPL environment. Pointcloud data (top row) is captured by the ground robots, and the color images (bottom row) are captured by the aerial robots. Each column shows one subgroup of one aerial robot and one ground robot exploring different parts of the environment.

at a height of 2.0~m above the ground robots, whereas the ground robots move at a speed of 0.5~m/s. The first team explores a lab room and the second one explores a hallway characterized by natural lighting.

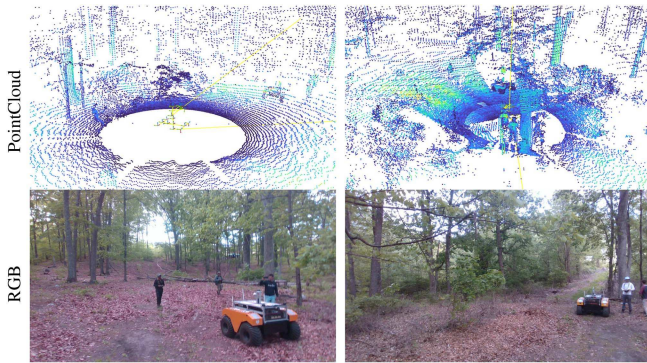


Fig. 4. Outdoor-Forest environment. Pointcloud data (top row) is captured by the ground robots, and the color images (bottom row) are captured by the aerial robots. This snapshot captures the scenarios when two groups of robot subteams encounter each other.

### C. Outdoor Sequences

In the outdoor scenario, we use two Warthog ground robots and two aerial robots. We present three sequences with two Warthog ground robots and two aerial robots, and two sequences with one Warthog ground robot and one aerial robot. Sample dataset qualitative results are shown in Fig. 4. The outdoor sequences are captured in two outdoor environments. The first environment, named ‘HOUSE’, is an open space of  $150\text{ m} \times 30\text{ m}$  situated by a forest, encompassing houses, roads, trees, and rocks. The houses in this setting can be used as object detection targets, thus providing a practical context for target detection tasks. The second environment, ‘FOREST’, is a space of  $100\text{ m} \times 80\text{ m}$  nestled within the forest, presenting a rich composition of trees, ground variations, and rocks. These environments introduce various features and obstacles commonly encountered in outdoor environments, thus offering a robust platform for training and testing algorithms. In all outdoor sequences, the robot team splits up into two subteams each composed of one aerial robot and one ground robot. The aerial robots fly above the ground robots for most of the time, and detach from the ground robot to explore the area where ground robots have difficulty maneuvering. The aerial robots fly at a height ranging from  $2.0\text{ m}$  to  $10.0\text{ m}$ . The ground robots drive at a speed up to  $1.5\text{ m/s}$ . In ‘Outdoor-HOUSEB’, we also experiment with switching the pairing of the aerial and ground robots in the middle of the sequence to study the effect of formation on collaborative perception performances.

## IV. CALIBRATION

*Aerial Robot Calibration:* The parameters of the aerial robot sensors are calibrated using the Kalibr toolbox [26]. The calibration process includes the intrinsic calibration [27] of the RGB and RGBD cameras respectively, the extrinsic calibration [28] of the RGBD camera and IMU, and the extrinsic calibration of the RGB camera and IMU. We also calibrate the IMU intrinsic parameter by conducting IMU Allan covariance analysis.



Fig. 5. Detected AprilTag mounted on the ground robots captured from one aerial robot’s downward-facing camera.

*Ground Robot Calibration:* We employ multi-sensor Graph-based Calibration to produce a unified graph-based representation of the ground robots’ sensor nodes [29]. These sensors include LiDAR, RGBD, and stereo cameras from Table II. The Ouster LiDAR (OS1) is chosen as the root/reference frame in the robot’s transform hierarchy. The graph optimization approach computes pairwise relative poses for each sensor node.

*Air-Ground Calibration:* To ensure the accuracy and consistency of our data, we implemented a meticulous air-ground system sensor calibration process. For extrinsic calibration of the start position of the robot team, a calibration board is positioned in front of all the forward-facing cameras of the robot team, which ensures the spatial alignment of different modalities and data consistency, therefore providing a reliable foundation for subsequent research.

## V. DATA ANNOTATION

### A. Pose Estimation

We employ an individual state estimation system for each robot. For aerial robots, the fusion of Global Positioning System (GPS) data with Stereo Visual Inertial Odometry provides accurate pose estimation. In our autonomy stack, we adopt OpenVINS [30] for stereo Visual Inertial Odometry (VIO). Conversely, the ground robots utilize IMU and wheel odometry for reliable state estimation. The maximum speed of the aerial robot is  $2.1\text{ m/s}$ , the maximum speed of the ground robot is  $1.75\text{ m/s}$ .

Moreover, computing the cross-robot relative pose estimation adds an additional level of precision. This is achieved by identifying a bundle of AprilTags [23]—fiducial marker systems commonly used in robotics—mounted on the ground robots from the vantage point of the aerial robots, illustrated in Fig. 5. We calibrate the extrinsic between each tag within the AprilTag bundle by capturing a sequence of images from a camera with known intrinsic parameters. Therefore, we can calculate the relative transformation between the AprilTag bundle frame with respect to the camera frame.

### B. Zero-Shot Semantics and Depth Annotation

In addition to pose estimation, we provide semantic and depth annotations within our dataset. We harness the power of foundation models to deliver zero-shot semantic and depth annotation. For semantics annotation, we adopt the combination of RAM [31], Grounding DINO [32] and SAM [33] to enable

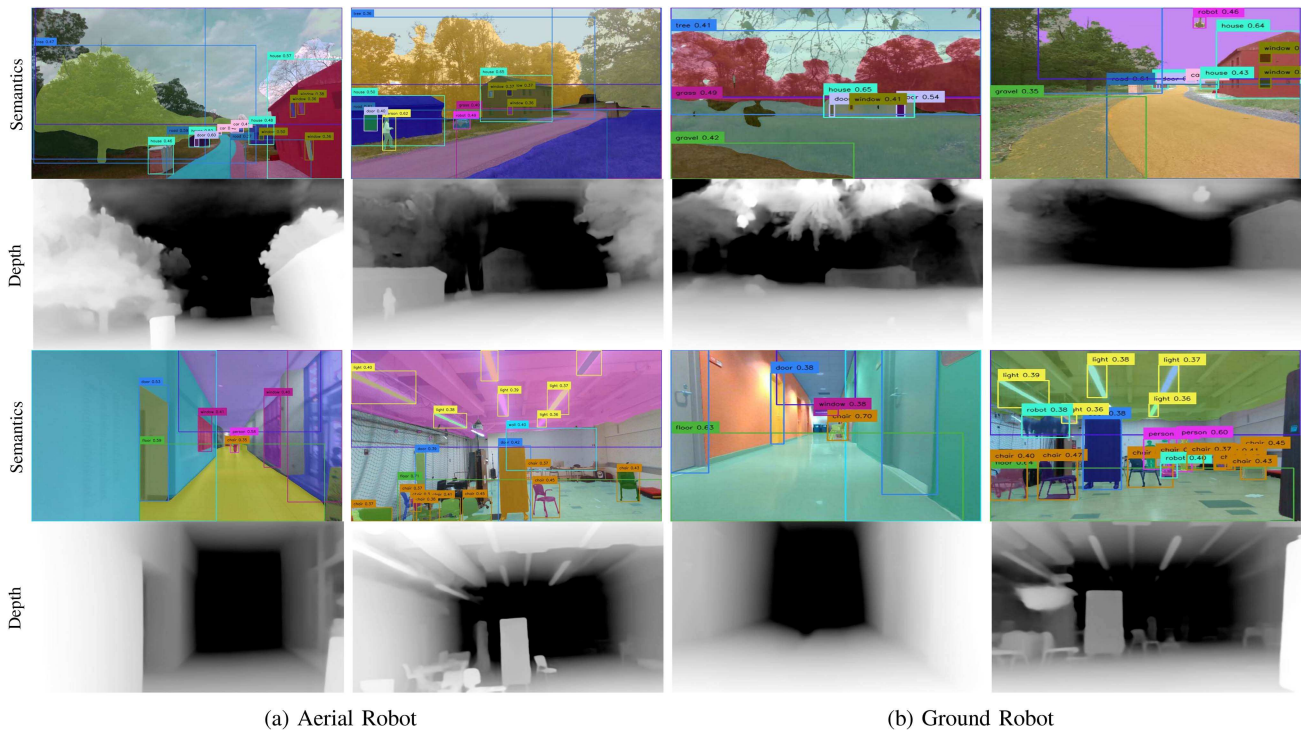


Fig. 6. Zero-shot annotation of semantics and depth in indoor and outdoor environments from aerial and ground robots.

the acquisition of 2D instance segmentation masks directly from the camera sensors. We obtain 2D bounding boxes by post-processing the mask annotations. For depth estimation, we adopt Zoedepth [34] to produce zero-shot monocular depth estimation annotation. Due to the limited baseline of Intel Realsense stereo camera, the quality of traditional stereo depth matching such as SGBM [35] and graph-cut stereo matching [36] are worse than state-of-the-art monocular depth estimation such as Zoedepth, which combines relative and metric depth. We cannot apply learning-based stereo-matching algorithms since these are designed for color rather than grayscale cameras which are employed on our robots. For semantics annotation, we utilize temporal propagation among multiple frames to increase the temporal consistency of the zero-shot prediction. Temporal consistent annotation is essential to denoise the annotation and it enables video-based dense prediction tasks. Specifically, we leverage [37] as temporal propagation model to propagate the segments on keyframes to consecutive frames. The employment of zero-shot learning, wherein the model is not explicitly trained on task-specific data, increases the generalization and applicability of the annotations. Therefore, they enhance the usability of our dataset, facilitating robust multi-robot collaborative perception research. Sample qualitative results are illustrated in Fig. 6.

## VI. DATASET ATTRIBUTES

The dataset presents several characteristics that are unique compared to other multi-robot datasets. These characteristics highlight real-world challenges and scenarios encountered when deploying a heterogeneous robot team in the wild. The statistics

TABLE IV  
SEQUENCE INFO

Sequence	Scene	Time	RGB Frames	Instances/Frame
NYUARPL	Indoor	400 s	48000	9
CHAIR-1	Indoor	180 s	10800	11
HALLWAY	Indoor	200 s	12000	3
HOUSEA	Outdoor	440 s	52800	8
HOUSEB	Outdoor	440 s	52800	9
FOREST	Outdoor	225 s	27000	5

of the scene is illustrated in Table IV. First, we present a mixed setup of indoor and outdoor scenarios. We showcase the transition between outdoor and indoor environments in ‘Outdoor-HOUSEB’ sequence using one aerial robot ‘race5’, this sequence is unique and challenging to perception algorithms, as it requires the robot to adapt to the drastic change in environment and sensor modalities. Second, for the multi-robot sequences, we provide continuous spatio-temporal variations of the formation and relative position of the robot team members, introducing distinct and complementary viewpoint perspectives to guarantee a wide range of inter-robot interactions. This variety is crucial in studying how different formations and positional arrangements influence collaborative perception and team efficiency. In the indoor sequence ‘Indoor-NYUARPL’ and outdoor sequence ‘Outdoor-FOREST’, we split the robot team in two groups, one aerial robot and one ground robot in each group. This is a challenging setup for perception algorithms, as it requires

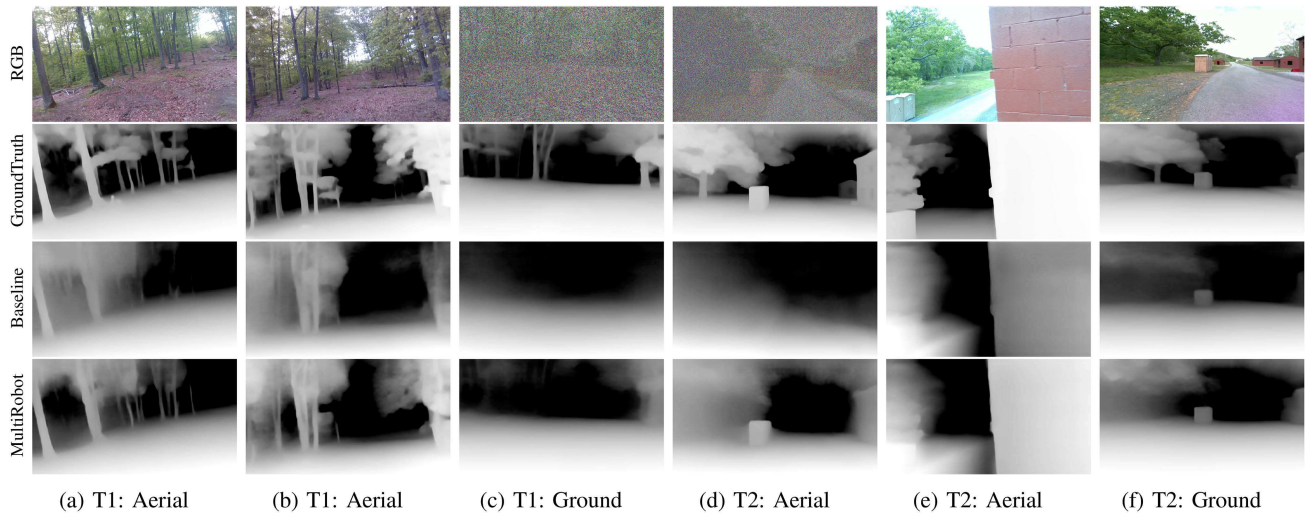


Fig. 7. Multi-robot collaborative depth estimation with two aerial robots and one ground robot. (a)–(c) and (d)–(f) show two scenarios (T1 and T2) with (c) and (d) robot image sensors corrupting the rows correspond to RGB images, groundtruth depth, single-robot baseline, and multi-robot collaboration results respectively.

the robot team to collaborate and share information to perform perception tasks under a wide range of overlap and occlusion conditions. Third, the dataset preserves the real-world sensor noise and disturbances encountered when deploying a heterogeneous robot team in the wild. Cameras may experience color jittering, LiDARs can exhibit discontinuous intensities temporally, GPS may be denied or drifted in indoor or outdoor forest environments, IMU may have inconsistent intrinsic parameters because of temperature change, and clock synchronization can be lost because of the network delay. These real-world disturbances are crucial for perception algorithms to be robust and reliable. Fourth, we also provide several single-robot sequences using one aerial robot exploring the ‘HOUSE’ and ‘FOREST’ environments. These sequences are captured using the same aerial robot as the multi-robot sequences, thus providing a consistent perspective for comparison. These sequences are useful to provide additional samples and annotations to facilitate perception research. Finally, the availability of scenes where aerial robots detach from the ground robots when these cannot maneuver due to environmental constraints further shows the diverse capabilities of the proposed heterogeneous system and its usefulness for collaborative perception tasks.

## VII. USE-CASES AND APPLICATIONS

Our dataset offers a unique platform for developing and evaluating collaborative perception algorithms in multi-robot systems. Researchers can explore the utilization of sensor data among various robots for complex multi-robot collaborative perception tasks including scenarios with sensor and environmental noise. We showcase qualitative results of monocular depth estimation and semantics segmentation using Graph Neural Network [12] in Figs. 7 and 8. Two aerial robots collaborate with one ground robot to overcome sensor noises by feature map communication. The results show that multi-robot collaboration improves the depth prediction and semantic segmentation

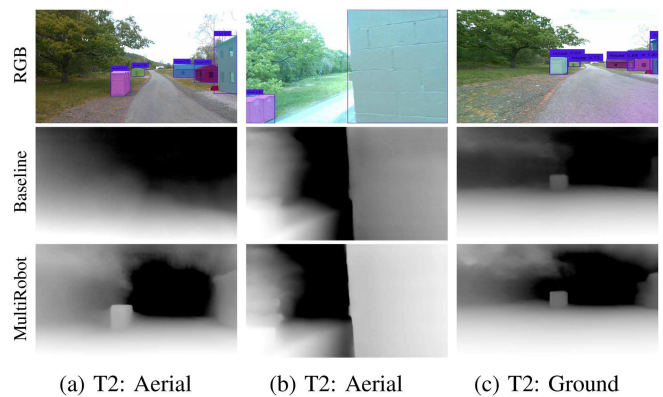


Fig. 8. Multi-robot collaborative semantics estimation with two aerial robots and one ground robot. Image sensor of (a) is corrupted. The rows correspond to groundtruth semantics, single-robot baseline, and multi-robot collaboration results.

robustness against single-agent baseline. The predictions of far-away objects are recovered, such as trees in depth estimation and houses in semantic segmentation.

Given that the dataset provides multi-modal sensor data from different types of robots, it can also be a tool for investigating sensor fusion techniques. This could lead to the development of more robust algorithms for multi-robot mapping [17], object detection [11], and decision-making [21] for real-world settings in varying indoor and outdoor conditions.

## VIII. CONCLUSION

In this letter, we presented a comprehensive multi-robot, multi-modal, and multi-rate air-ground dataset captured in diverse real-world indoor and outdoor environments. The goal is to provide a substantial resource for the research community to investigate and advance the field of multi-robot collaborative perception. Specifically, the dataset includes data from two

types of robots (i.e., ground and aerial) equipped with multiple heterogeneous sensing modalities. A substantial effort was invested in covering multiple distinct indoor and outdoor environments, specifically 'HOUSE' and 'FOREST'. This involved varying navigation sequences and formation therefore ensuring to capture heterogeneous sensing data affected by real-world sensor noise and disturbances and with large spatio-temporal distinct viewpoints variations and complementary perspectives. The data and annotations offer a unique opportunity to delve into several challenging research problems and aspects related to multi-robot systems, such as collaborative perception, sensor fusion, object detection, and formation control. We believe that the limitations and challenges associated with our dataset will be a source of inspiration to the community to advance the research development of multi-robot perception systems and autonomy algorithms.

In the future, we plan to enrich the dataset with comprehensive 2D and 3D semantic annotations. This dataset should not be considered a monolithic solution. We will continue to expand it, drawing from additional tests in multiple additional environmental contexts and with diverse robot configurations. Feedback and suggestions from the community will be pivotal in guiding these enhancements. Finally, this dataset can serve to improve robotics simulation by real-to-sim. It will serve as a stepping stone for future research in collaborative perception and will be instrumental in shaping the future of autonomous multi-robot systems.

#### ACKNOWLEDGMENT

The authors would like to thank Julie Foresta, Nishanth Bobbili, Roshan Balu T. M. B., Luca Morando, Jeffery Mao, and Guanrui Li for their help during the dataset collection process.

#### REFERENCES

- [1] Y. Chang et al., "LAMP 2.0: A robust multi-robot SLAM system for operation in challenging large-scale underground environments," *IEEE Robot. Automat.*, vol. 7, no. 4, pp. 9175–9182, Oct. 2022.
- [2] M. Tranzatto et al., "Team cerberus wins the darpa subterranean challenge: Technical overview and lessons learned," 2022, *arXiv:2207.04914*.
- [3] Y. Zhu, Y. Kong, Y. Jie, S. Xu, and H. Cheng, "GRACO: A multimodal dataset for ground and aerial cooperative localization and mapping," *IEEE Robot. Automat. Lett.*, vol. 8, no. 2, pp. 966–973, Feb. 2023.
- [4] D. Feng et al., "S3E: A large-scale multimodal dataset for collaborative SLAM," 2022, *arXiv:2210.13723*.
- [5] Y. Tian et al., "Resilient and distributed multi-robot visual SLAM: Datasets, experiments, and lessons learned," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 11027–11034.
- [6] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 2583–2589.
- [7] E. Arnold, M. Dianati, R. d. Temple, and S. Fallah, "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, Mar. 2022.
- [8] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21329–21338.
- [9] Y. Li et al., "V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022.
- [10] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4105–4114.
- [11] Y. Hu, S. F. Fang, Z. Lei, Y. Z. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. Thirty-sixth Conf. Neural Inf. Process. Syst.*, 2022, pp. 4874–4886.
- [12] Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno, "Multi-robot collaborative perception with graph neural networks," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2289–2296, Apr. 2022.
- [13] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng, "Multi-robot scene completion: Towards task-agnostic collaborative perception," in *Proc. Conf. Robot Learn.*, 2023, pp. 2062–2072.
- [14] P.-Y. Lajoie, B. Ramtoula, F. Wu, and G. Beltrame, "Towards collaborative simultaneous localization and mapping: A survey of the current research landscape," *Field Robot.*, vol. 2, pp. 971–1000, 2022.
- [15] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6876–6883.
- [16] C. Fan, J. Hu, and J. Huang, "Few-shot multi-agent perception," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1712–1720.
- [17] P. Gao, Q. Zhu, H. Lu, C. Gan, and H. Zhang, "Deep masked graph matching for correspondence identification in collaborative perception," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 6117–6123.
- [18] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 29541–29552.
- [19] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 6035–6042.
- [20] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 316–332.
- [21] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17231–17241.
- [22] M. Quigley et al., "ROS: An open-source robot operating system," in *Proc. ICRA Workshop Open Source Softw.*, 2009, vol. 3, no. 3, p. 5.
- [23] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3400–3407.
- [24] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 252–286, 2014.
- [25] D. Malyuta, "Guidance, navigation, control and mission logic for quadrotor full-cycle autonomy," Master's thesis, ETH Zurich, Zurich, Switzerland, 2018.
- [26] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 4304–4311.
- [27] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
- [28] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1280–1286.
- [29] J. L. Owens, P. R. Osteen, and M. Daniilidis, "MSG-cal: Multi-sensor graph-based calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 3660–3667.
- [30] P. Geneva, K. Eickenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 4666–4672.
- [31] Y. Zhang et al., "Recognize anything: A strong image tagging model," 2023, *arXiv:2306.03514*.
- [32] S. Liu et al., "Grounding Dino: Marrying Dino with grounded pre-training for open-set object detection," 2023, *arXiv:2303.05499*.
- [33] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.
- [34] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," 2023, *arXiv:2302.12288*.
- [35] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 807–814.
- [36] V. Kolmogorov, P. Monasse, and P. Tan, "Kolmogorov and Zabih's graph cuts stereo matching algorithm," *Image Process. Line*, vol. 4, pp. 220–251, 2014.
- [37] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," 2023, *arXiv:2304.11968*.