# C. RoboSense: Large-scale Dataset and Benchmark for Egocentric Robot Perception and Navigation in Crowded and Unstructured Environments

# C. RoboSense:大规模数据集与基准，用于拥挤和非结构化环境中的自我中心机器人感知与导航

Haisheng Feixiang Cong Wei Junchi School of Computer Science, Shanghai Jiao Tong University SenseAuto Research

海胜 飞翔 聪 伟 俊驰 上海交通大学计算机科学与工程学院 SenseAuto研究院

{suhaisheng, yanjunchi}@sjtu.edu.cn, {songfeixiang1, macong, wuwei}@senseauto.com Code & Dataset: SenseAuto & SJTU-ReThinkLab/RoboSense

{suhaisheng, yanjunchi}@sjtu.edu.cn, {songfeixiang1, macong, wuwei}@senseauto.com 代码与数据集:SenseAuto & SJTU-ReThinkLab/RoboSense

# Abstract

# 摘要

Reliable embodied perception from an egocentric perspective is challenging yet essential for autonomous navigation technology of intelligent mobile agents. With the growing demand of social robotics, near-field scene understanding becomes an important research topic in the areas of egocentric perceptual tasks related to navigation in both crowded and unstructured environments. Due to the complexity of environmental conditions and difficulty of surrounding obstacles owing to truncation and occlusion, the perception capability under this circumstance is still inferior. To further enhance the intelligence of mobile robots, in this paper, we setup an egocentric multi-sensor data collection platform based on 3 main types of sensors (Camera, LiDAR and Fisheye), which supports flexible sensor configurations to enable dynamic sight of view from ego-perspective, capturing either near or farther areas. Meanwhile, a large-scale multimodal dataset is constructed, named RoboSense, to facilitate egocentric robot perception. Specifically, RoboSense contains more than synchronized data with bounding box and IDs annotated in the full view, forming trajectories across temporal sequences. It has and as many annotations of surrounding obstacles within near ranges as the previous datasets collected for autonomous driving scenarios such as KITTI and nuScenes. Moreover, we define a novel matching criterion for near-field 3D perception and prediction metrics. Based on RoboSense, we formulate 6 popular tasks to facilitate the future research development, where the detailed analysis as well as benchmarks are also provided accordingly. Data desensitization measures have been conducted for privacy protection.

从自我中心视角进行可靠的具身感知具有挑战性，但对于智能移动代理的自主导航技术至关重要。随着社交机器人需求的增长，近场场景理解成为与拥挤和非结构化环境中导航相关的自我中心感知任务领域的重要研究课题。由于环境条件的复杂性以及截断和遮挡导致的周围障碍物的难度，这种情况下的感知能力仍然不足。为了进一步增强移动机器人的智能，本文中我们建立了一个基于三种主要传感器(相机、激光雷达和鱼眼镜头)的自我中心多传感器数据采集平台，支持灵活的传感器配置，以实现从自我视角的动态视野，捕捉近处或更远的区域。同时，构建了一个名为RoboSense的大规模多模态数据集，以促进自我中心机器人感知。具体而言，RoboSense包含超过 的同步数据，其中 个边界框和ID在全 视图中标注，形成 条轨迹，跨越 个时间序列。它在近范围内的周围障碍物标注数量是之前为自动驾驶场景收集的数据集(如KITTI和nuScenes)的 和 倍。此外，我们定义了近场3D感知和预测指标的新匹配标准。基于RoboSense，我们制定了6个流行任务，以促进未来的研究发展，并提供了详细的分析和基准。已采取数据脱敏措施以保护隐私。

# 1. Introduction

# 1. 引言

Recent years have witnessed significant progress achieved in the field of autonomous driving, enabling numerous intelligent vehicles running on highway or urban areas. In addition to self-driving cars, social mobile robots have emerged as a new industry tailored to autonomous navigation for typical applications, such as tractor, sweeper, retail and delivery. Notably, such intelligent mobile agents usually operate and navigate in crowded and unstructured environments (i.e., campuses, scenic spots, streets, parks and sidewalks, etc.), with varying and uncontrolled natural conditions such as illumination, occlusion and obstruction. In order to achieve navigation tasks safely, egocentric perceptual solutions enable these robots to perceive and comprehend the surrounding context from a first-person view, so as to interact successfully with passby pedestrians and vehicles, predict their intentions and incorporate this information in agents’ planning and decision reasoning process.

近年来，自动驾驶领域取得了显著进展，使得众多智能车辆能够在高速公路或城市区域运行。除了自动驾驶汽车外，社交移动机器人已成为一个新兴行业，专为典型应用的自主导航而设计，如拖拉机、清扫车、零售和配送。值得注意的是，这些智能移动代理通常在拥挤和非结构化环境(如校园、景点、街道、公园和人行道等)中操作和导航，环境条件多变且不受控制，如光照、遮挡和阻碍。为了安全地完成导航任务，自我中心感知解决方案使这些机器人能够从第一人称视角感知和理解周围环境，从而成功与过往行人和车辆互动，预测他们的意图，并将这些信息纳入代理的规划和决策推理过程中。

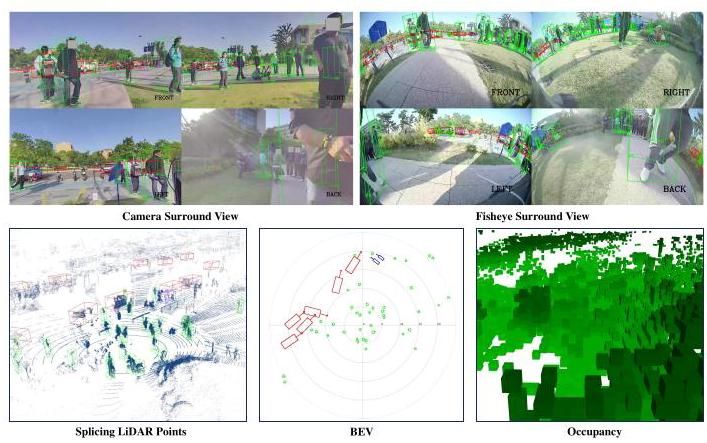


Figure 1. An example from RoboSense dataset: The data with annotated 3D boxes and occupancy descriptions on Camera, Fisheye, LiDAR, and BEV respectively, where the same targets are associated with unique IDs across different devices and timestamps.

图1. RoboSense数据集中的一个示例:分别在相机、鱼眼镜头、激光雷达和BEV上标注了3D框和占用描述的数据，其中相同目标在不同设备和时间戳上关联了唯一ID。

To evaluate and compare different egocentric perceptual methods fairly, several standarized benchmarks , have been proposed in recent years, advancing the development of modern data-driven approaches. KITTI [8] is a pioneering dataset providing multi-modal sensor data including front-view LiDAR pointclouds as well as corresponding stereo images and GPS / IMU data. nuScenes [4] constructs a multi-sensor dataset collected in two cities travelling at an average of , where rich collections of boxes and IDs are annotated in the full view. Waymo Open dataset [33] significantly increases the amount of annotations with higher annotation frequency. However, the target domain application of existing benchmarks is autonomous driving: the sensor data are captured exclusively from structural roads and highways, with sensor suites installed on top of cars.

为了公平评估和比较不同的自我中心感知方法，近年来提出了几个标准化基准 、 ，推动了现代数据驱动方法的发展。KITTI [8] 是一个开创性的数据集，提供了多模态传感器数据，包括前视激光雷达点云以及相应的立体图像和GPS/IMU数据。nuScenes [4] 构建了一个在两个城市中以平均 速度行驶时收集的多传感器数据集，其中在完整的 视图中标注了丰富的 框和ID。Waymo开放数据集 [33] 显著增加了标注数量，并提高了标注频率。然而，现有基准的目标应用是自动驾驶:传感器数据仅从结构化道路和高速公路上捕获，传感器套件安装在汽车顶部。

To fill the vacancies of egocentric perceptual benchmarks target a unique domain related to navigation tasks in crowded and unstructured environments, in this paper, we present RoboSense, a novel multimodal dataset with several benchmarks associated to it. Our dataset is collected from diverse social scenarios filled with crowded obstructions, which is different from previously collected datasets used for autonomous driving (e.g. nuScenes [4]). Benefiting from the well time-synced multi-sensor data, we hope that our RoboSense can facilitate the development of egocentric perceptual frameworks for various types of autonomous navigation agents with controllable cost, not only self-driving cars but also autonomous agents such as social mobile robots. To this end, the data collection robot is equipped with 3 main types of sensors (C: Camera, L: LiDAR, F: Fisheye), and each type of sensor consists of 4 devices installed on different sides respectively to ensure the data captured under full view without blind spots.

为了填补自我中心感知基准在拥挤和非结构化环境中与导航任务相关的独特领域的空白，本文提出了RoboSense，这是一个新颖的多模态数据集，并附有几个相关基准。我们的数据集从充满拥挤障碍物的多样化社交场景中收集，与之前用于自动驾驶的数据集(如nuScenes [4])不同。得益于良好时间同步的多传感器数据，我们希望RoboSense能够促进各种类型自主导航代理的自我中心感知框架的发展，不仅限于自动驾驶汽车，还包括社交移动机器人等自主代理。为此，数据收集机器人配备了3种主要类型的传感器(C:摄像头，L:激光雷达，F:鱼眼镜头)，每种类型的传感器分别由4个设备组成，安装在不同的侧面，以确保在完整的 视图中捕获数据，无盲区。

Specifically, RoboSense consists of a total of frames of synchronized data, spanning over temporal sequences of 6 main scene classes (i.e., scenic spots, parks, squares, campuses, streets and sidewalks). Moreover, 1.4M 3D bounding boxes together with track IDs are annotated based on 3 different types of sensors, where most of targets tend to be closer to the robot as shown in Fig. 1. Then we form global trajectories for each agent separately through associating the same IDs across consecutive frames and different devices from a Bird’s-Eye View (BEV) perspective. Additionally, we formulate 6 standarized benchmarks for egocentric perceptual tasks as follows: 1. Multi-view 3D Detection; 2. LiDAR 3D Detection; 3. Multi-modal 3D Detection; 4. Multiple 3D Object Tracking (3D MOT); 5. Motion Prediction; 6. Occupancy Prediction. Meanwhile, multi-task end-to-end training scheme is also supported in our RoboSense for evaluation of joint optimization. In sum, the main contributions of our work are three folds:

具体来说，RoboSense由总共 帧同步数据组成，跨越 个时间序列，涵盖6个主要场景类别(即景点、公园、广场、校园、街道和人行道)。此外，基于3种不同类型的传感器标注了140万个3D边界框和跟踪ID，其中大多数目标倾向于更接近机器人，如图1所示。然后，我们通过从鸟瞰图(BEV)视角关联连续帧和不同设备中的相同ID，为每个代理分别形成全局轨迹。此外，我们为自我中心感知任务制定了6个标准化基准:1. 多视图3D检测；2. 激光雷达3D检测；3. 多模态3D检测；4. 多目标3D跟踪(3D MOT)；5. 运动预测；6. 占用预测。同时，我们的RoboSense还支持多任务端到端训练方案，用于评估联合优化。总之，我们工作的主要贡献有三点:

* To our best knowledge, our RoboSense is the first dataset tailored to egocentric perceptual tasks related to navigation of autonomous agents in unstructured environments.
* 据我们所知，RoboSense是第一个专门为非结构化环境中自主代理导航相关的自我中心感知任务量身定制的数据集。
* We annotate bounding boxes on synchronized sensor data, where most of targets are closer to the robot. Each target is associated with a unique ID, thus forming a total of trajectories, which spread over temporal sequences, covering 6 main scene classes.
* 我们在 同步传感器数据上标注了 个边界框，其中大多数目标更接近机器人。每个目标都与一个唯一的ID相关联，从而形成了总共 条轨迹，这些轨迹分布在 个时间序列中，覆盖了6个主要场景类别。
* We formulate 6 standardized benchmarks to facilitate the evaluation and fair comparisons of different perceptual solutions related to navigation in built environments.
* 我们制定了6个标准化基准，以促进与建筑环境中导航相关的不同感知解决方案的评估和公平比较。

# 2. Related Work

# 2. 相关工作

We summarize the compositions of some existing perception and prediction datasets as shown in Tab. 1.

我们总结了一些现有感知和预测数据集的组成，如表1所示。

Perception Datasets. Current released perception datasets can be divided into image-only datasets [6, 42] and multimodal datasets [4, 8, 15, 16, 33]. BDD100k [42] and Cityscapes [6] focus on 2D perception which provide large amount of annotations (boxes, masks) for driving scene understanding under various weather and illumination conditions. KITTI [8] is known as the pioneering multimodal dataset which has been widely used for academic research. It records 6 hours of driving data using a LiDAR sensor and a front-facing stereo camera to provide pointclouds and images with annotated 3D boxes. H3D dataset [25] collects a total of objects over frames from 160 crowded scenes of the full view. nuScenes [4] and Waymo Open Dataset [33] are two similar datasets with same structure, while the latter one providing more annotations owing to higher annotation frequency vs. ). Different from previously collected datasets used for autonomous driving, the annotation frequency of our Ro-boSense is even smaller(1Hz)due to the low speed (less than ) moving status of social mobile robots navigating in crowded and unstructured environments.

感知数据集。当前发布的感知数据集可分为仅图像数据集[6, 42]和多模态数据集[4, 8, 15, 16, 33]。BDD100k [42]和Cityscapes [6]专注于2D感知，提供了大量 标注(框、掩码)用于各种天气和光照条件下的驾驶场景理解。KITTI [8]是众所周知的开创性多模态数据集，已广泛用于学术研究。它使用LiDAR传感器和前置立体相机记录了6小时的驾驶数据，提供带有标注3D框的点云和图像。H3D数据集[25]从160个拥挤场景的全 视图中收集了总共 个对象，分布在 帧中。nuScenes [4]和Waymo Open Dataset [33]是两个结构相似的数据集，后者由于更高的标注频率 vs. 提供了更多标注。与之前用于自动驾驶的数据集不同，我们的Ro-boSense的标注频率更低(1Hz)，因为社交移动机器人在拥挤和非结构化环境中导航时的移动速度较慢(小于 )。

Prediction Datasets. nuScenes [4] and Waymo Open Dataset [33] can be also used for prediction task which release lane graphs as well. Lyft [16] introduces traffic/speed control data, and Waymo Open Dataset [33] adds more signals to the map such as crosswalk, lane boundaries, stop signs and speed limits. Recently, Shifts dataset [24] becomes the largest forecasting dataset with the most scenario hours to date. Meanwhile, Argoverse [5] is also a large-scale dataset with high data frequency(10Hz)and high scenario quality for motion forecasting across 6 cities). Together, these datasets have enabled exploration of multi-actor, long-range motion forecasting leveraging both static and dynamic maps.

预测数据集。nuScenes [4]和Waymo Open Dataset [33]也可用于预测任务，它们还发布了车道图。Lyft [16]引入了交通/速度控制数据，Waymo Open Dataset [33]在地图中添加了更多信号，如人行横道、车道边界、停车标志和速度限制。最近，Shifts数据集[24]成为迄今为止最大的预测数据集，拥有最多的场景小时数。同时，Argoverse [5]也是一个大规模数据集，具有高数据频率(10Hz)和高场景质量，用于跨6个城市的运动预测 。这些数据集共同支持了利用静态和动态地图进行多参与者、长距离运动预测的探索。

Generally, our dataset differs in three substantial ways: 1) targets a unique domain related to navigation tasks in crowded and unstructured environments, which is more difficult than autonomous driving scenarios in terms of complexity of environmental context and diversity of surrounding obstructions. 2) In addition to 3D bounding box and trajectory annotations, our dataset also provides high-quality occupancy descriptions for each collected scene, supporting the occupancy prediction task around the social robotics for safe navigation. 3) Our dataset is mostly collected in social crowded scenes, where pedestrians and cars tend to be closer to the robot, yielding a distribution with a mode at approximately , which is quite different to the existing datasets for autonomous cars as shown in Fig. 2. Besides, the egocentric perceptual tasks under this circumstance is more challenging due to frequent occlusion and truncation.

总的来说，我们的数据集在三个重要方面有所不同:1)针对拥挤和非结构化环境中的导航任务，这一领域在环境背景的复杂性和周围障碍物的多样性方面比自动驾驶场景更具挑战性。2)除了3D边界框和轨迹标注外，我们的数据集还提供了每个收集场景的高质量占用描述，支持社交机器人周围的安全导航占用预测任务。3)我们的数据集主要在社交拥挤场景中收集，行人和汽车往往更接近机器人，产生的分布模式大约在 ，这与现有的自动驾驶汽车数据集有很大不同，如图2所示。此外，在这种情况下，以自我为中心的感知任务由于频繁的遮挡和截断而更具挑战性。

Table 1. Statistical comparison between RoboSense and similar existing datasets used for autonomous driving. C: Camera, L: LiDAR, F: Fisheye. means statistics exclude the testing set, which is unavailable. indicates higher annotation frequency (10Hz).

表1. RoboSense与用于自动驾驶的类似现有数据集之间的统计比较。C:相机，L:LiDAR，F:鱼眼。 表示统计数据不包括测试集，因为测试集不可用。 表示 更高的标注频率(10Hz)。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Year | Size (hr) | Scenes | Frames | With Trajectory | Multi-view Overlapping | Sensor Layouts | 3D Boxes ( Total ) | 3D Boxes |
| KITTI [8] | 2012 | 1.5 | 22 | 15K | ✘ | ✘ | 4C+1L | 80K | 638 |
| Cityscapes [6] | 2016 | - | - | 25K | ✘ | ✘ | 1C | 0 | 0 |
| ApolloScape [15] | 2016 | 2 | - | 144K | ✘ | ✘ | 1L | 70K | 4.7K |
| H3D [25] | 2019 | 0.77 | 160 | 27K | ✘ | ✓ | 3C+1L | 1.1M | - |
| Lyft L5 [16] | 2019 | 2.6 | 366 | 55K | ✓ | ✓ | 7C+3L | 1.3M | - |
| nuScenes [4] | 2019 | 5.5 | 1K | 40K | ✓ | ✓ | 6C+1L | 1.4M | 9.8K |
| Argoverse [5] | 2019 | 0.6 | 113 | 22K | ✓ | ✓ | 9C+2L | 993K | 15K |
| Waymo Open [33] | 2019 | 6.4 | 1K | 200K‡ | ✓ | ✓ | 5C+5L | 12Mt | 123Kt |
| BDD100k [42] | 2020 | 1K | 100k | 100k | ✘ | ✘ | 1C | 0 | 0 |
| RoboSense (Ours) | 2024 | 42 | 7.6K | 133K | ✓ | ✓ | 4C+4F+4L | 1.4M | 173K |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 年份 | 大小(小时) | 场景 | 帧 | 带有轨迹 | 多视图重叠 | 传感器布局 | 3D 框(总计) | 3D 框 |
| KITTI [8] | 2012 | 1.5 | 22 | 15K | ✘ | ✘ | 4C+1L | 80K | 638 |
| Cityscapes [6] | 2016 | - | - | 25K | ✘ | ✘ | 1C | 0 | 0 |
| ApolloScape [15] | 2016 | 2 | - | 144K | ✘ | ✘ | 1L | 70K | 4.7K |
| H3D [25] | 2019 | 0.77 | 160 | 27K | ✘ | ✓ | 3C+1L | 1.1M | - |
| Lyft L5 [16] | 2019 | 2.6 | 366 | 55K | ✓ | ✓ | 7C+3L | 1.3M | - |
| nuScenes [4] | 2019 | 5.5 | 1K | 40K | ✓ | ✓ | 6C+1L | 1.4M | 9.8K |
| Argoverse [5] | 2019 | 0.6 | 113 | 22K | ✓ | ✓ | 9C+2L | 993K | 15K |
| Waymo Open [33] | 2019 | 6.4 | 1K | 200K‡ | ✓ | ✓ | 5C+5L | 12Mt | 123Kt |
| BDD100k [42] | 2020 | 1K | 100k | 100k | ✘ | ✘ | 1C | 0 | 0 |
| RoboSense(我们的) | 2024 | 42 | 7.6K | 133K | ✓ | ✓ | 4C+4F+4L | 1.4M | 173K |

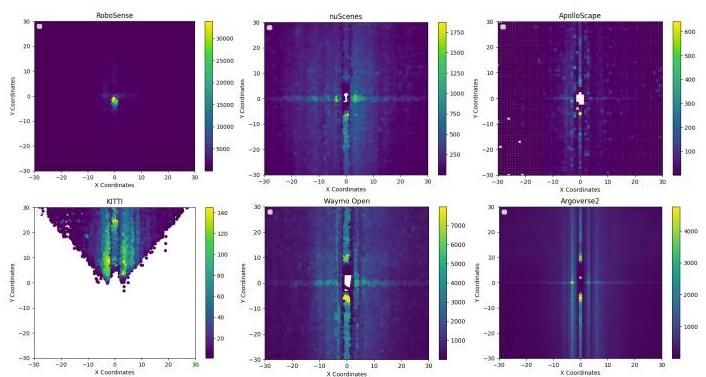


Figure 2. Comparison of annotated object distribution among different popular datasets used for perception and prediction tasks.

图2. 用于感知和预测任务的不同流行数据集中标注对象分布的对比。

# 3. RoboSense Open Dataset

# 3. RoboSense开放数据集

We commence with the sensor setup as well as data acquisition details, delineate the coordinate systems and label generation process, and present data statistics respectively.

我们从传感器设置以及数据采集细节开始，描述坐标系和标签生成过程，并分别展示数据统计。

# 3.1. Sensor Setup and Data Acquisition

# 3.1. 传感器设置与数据采集

Sensor setup. We use a social mobile robot (i.e., ro-bosweeper) as data collection platform, which is equipped with different sensors installed in different sides of the robot respectively to ensure data captured in horizontal view without blind spots, including LiDAR, Camera, Fisheye, GPS / IMU and Ultrasonic. Refer to Fig. 3 for sensor layouts and Tab. 3 for detailed sensor specifications.

传感器设置。我们使用一台社交移动机器人(即ro-bosweeper)作为数据采集平台，该机器人配备了分别安装在机器人不同侧面的多种传感器，以确保在 水平视角下无盲区地捕获数据，包括激光雷达(LiDAR)、摄像头、鱼眼镜头、GPS/IMU和超声波传感器。传感器布局请参见图3，详细传感器规格请参见表3。

Data acquisition. We utilize the mobile robot to collect data along the Dishui Lake in Shanghai, China, lasting 42h in total at an average speed of less than through manually remote control. 22 different places are travelled, which can be categorized into 6 main kinds of outdoor or semiclosed social scenarios (i.e., scenic spots, parks, squares, campuses, streets and sidewalks). After data collection, we manually select and process 7619 representative scenes of duration respectively for further annotation, covering various natural conditions (i.e., weather and illumination) and diverse environmental background and obstructions (i.e., motion, amount, type, occlusion, truncation).

数据采集。我们利用移动机器人在中国上海滴水湖沿线采集数据，总时长为42小时，平均速度低于 ，通过手动遥控操作。共经过22个不同地点，这些地点可分为6种主要的户外或半封闭社交场景(即景点、公园、广场、校园、街道和人行道)。数据采集后，我们手动选择并处理了7619个具有代表性的场景，每个场景的时长为 ，涵盖各种自然条件(即天气和光照)以及多样化的环境背景和障碍物(即运动、数量、类型、遮挡、截断)。

# 3.2. Coordinate Systems

# 3.2. 坐标系

Ego-Vehicle Coordinate. The Ego-Vehicle Coordinate System is centered at the rear axle of the vehicle. The positive directions of the , and axes correspond to the forward, leftward, and upward directions of the vehicle, respectively. Ego-Vehicle Coordinate System is the most frequently used in tasks such as perception, tracking, prediction, and planning, where dynamic and static targets as well as trajectories are transformed into this coordinate system.

自车坐标系。自车坐标系以车辆后轴为中心。 、 轴的正方向分别对应车辆的前方、左方和上方。自车坐标系在感知、跟踪、预测和规划等任务中最常用，动态和静态目标以及轨迹都会转换到该坐标系中。

Global Coordinate. To transform the dynamic and static elements from historical and future frames into the current frame coordinate system, we need to establish a global coordinate system to record the position and orientation of the ego vehicle in each frame. The origin of the Global Coordinate System is an arbitrarily defined point in Shanghai Lin-gang, China, and the positive directions of the , and axes follow the definition of the North-East-Up coordinate. LiDAR Coordinate. The LiDAR Coordinate System is defined based on the Hesai lidar installed directly above the vehicle, the positive directions of the , and axes follow the definition of the Ego-Vehicle Coordinate System.

全局坐标系。为了将历史和未来帧中的动态和静态元素转换到当前帧坐标系中，我们需要建立一个全局坐标系来记录自车在每一帧中的位置和方向。全局坐标系的原点是中国上海临港的一个任意定义点， 、 轴的正方向遵循东北天坐标系的定义。激光雷达坐标系。激光雷达坐标系基于安装在车辆正上方的禾赛激光雷达定义， 、 轴的正方向遵循自车坐标系的定义。

Camera Coordinate. The RoboSweeper is equipped with four fisheye cameras and four pinhole cameras. The origin of the Camera Coordinate System for both types of cameras is the optical center. However, the positive directions of the coordinate axes are defined differently in the RoboSense dataset. In the fisheye coordinate system, the X, Y, and Z axes correspond to directly below, right, and behind the optical center, respectively. In contrast, in the pinhole coordinate system, these axes correspond to directly right, below, and front of the optical center, respectively.

摄像头坐标系。RoboSweeper配备了四个鱼眼摄像头和四个针孔摄像头。两种摄像头的摄像头坐标系原点均为光学中心。然而，在RoboSense数据集中，坐标轴的正方向定义不同。在鱼眼坐标系中，X、Y、Z轴分别对应光学中心的正下方、右方和后方。而在针孔坐标系中，这些轴分别对应光学中心的正右方、下方和前方。

Pixel Coordinate. The image is presented in the form of pixels, each pixel corresponds to a 2D pixel coordinate. The origin of the Pixel Coordinate System is the upper left corner of the image. Points in the 3D Camera Coordinate System can obtain coordinates in the Pixel Coordinate System through the camera projection.

像素坐标系。图像以像素形式呈现，每个像素对应一个二维像素坐标。像素坐标系的原点是图像的左上角。三维摄像头坐标系中的点可以通过摄像头投影获得像素坐标系中的坐标。



Figure 3. Sensor setup and coordinate system illustration of our data collection platform.

图3. 我们数据采集平台的传感器设置和坐标系示意图。

# 3.3. Ground Truth Labels

# 3.3. 真实标签

After integrating, synchronizing and calibrating the multi-sensor raw data, we annotate keyframes (LiDAR, image) at the frequency of due to the low-speed moving status.

在整合、同步和校准多传感器原始数据后，由于低速移动状态，我们以 的频率对关键帧(激光雷达、图像)进行标注。

3D object. With the selected scenes of collected RoboSense dataset, we annotate 3D object boxes of 3 movable classes (i.e., "Vehicle", "Cyclist" and "Pedestrian") for each sampled keyframe in both the LiDAR coordinate of point-clouds and the Camera coordinate of multi-view images respectively. Each annotated 3D box can be represented as , where indicate the 3D position of a regular object, and represent the scale information including width, length and height. and correspond to the orientation (especially yaw angle) and the object class respectively. A three-stage auto-labelling pipeline is detailed in the supplementary material (see Sec. B.2).

三维物体。利用收集的RoboSense数据集中的选定场景，我们分别在点云的LiDAR坐标系和多视角图像的相机坐标系中为每个采样的关键帧标注了3个可移动类别(即“车辆”、“骑行者”和“行人”)的三维物体框。每个标注的三维框可以表示为 ，其中 表示规则物体的三维位置， 表示包括宽度、长度和高度的尺度信息。 和 分别对应方向(特别是偏航角)和物体类别。三阶段自动标注流程在补充材料中有详细说明(见第B.2节)。

Trajectory. To facilitate the temporal tasks such as multi-object tracking and motion forecasting described in Sec. 4, we assign a unique Track ID to each agent across a temporal sequence on Bird-Eye-View (BEV) of the Ego-Vehicle coordinate. Furthermore, agents with the same within a sequence are linked together to form object trajectories.

轨迹。为了促进第4节中描述的多目标跟踪和运动预测等时间任务，我们在自车坐标系的鸟瞰图(BEV)上为每个代理分配了一个唯一的轨迹ID 。此外，序列中具有相同 的代理被链接在一起，形成物体轨迹。

Occupancy label. In addition to 3 typical classes of moving objects on roads which are annotated temporally as above, there also exists a rich collection of static obstacles with irregular shapes especially in the complex scenarios (i.e., parks, campuses and squares, etc.) of RoboSense. To detailly describe the environment in surrounding camera views for driving safety, we voxelize the 3D space and generate high-quality yet dense occupancy labels to represent the voxel states. Similar with previous occupancy benchmarks [35, 36] built upon public datasets [4, 33], we conduct dynamic objects and static scenes segmentation along the temporal dimension based on annotated boxes and trajectories. Then sparse LiDAR points inside each box are extracted from to frames respectively, where indicates the index of current keyframe, and is set to 10 empirically. Refer to the supplementary material for more details of occupancy label generation process (see Sec. B.3).

占据标签。除了上述按时间标注的道路上三类典型移动物体外，RoboSense中还存在大量形状不规则的静态障碍物，尤其是在复杂场景(如公园、校园和广场等)中。为了详细描述周围相机视图中的环境以确保驾驶安全，我们对三维空间进行体素化，并生成高质量且密集的占据标签来表示体素状态。与之前基于公共数据集[4, 33]构建的占据基准[35, 36]类似，我们基于标注的 框和轨迹沿时间维度进行动态物体和静态场景的分割。然后分别从 到 帧中提取每个框内的稀疏LiDAR点，其中 表示当前关键帧的索引， 根据经验设置为10。有关占据标签生成过程的更多细节，请参阅补充材料(见第B.3节)。

# 4. Tasks & Metrics

# 4. 任务与指标

Both egocentric perceptual tasks and prediction tasks are supported in our RoboSense dataset and benchmark.

我们的RoboSense数据集和基准支持自中心感知任务和预测任务。

# 4.1. Perception

# 4.1. 感知

# 4.1.1 3D Object Detection

# 4.1.1 三维物体检测

The RoboSense 3D detection task requires to detect bounding boxes of three main classes (i.e. "Vehicle", "Pedestrian" and "Cyclist"), including position, size, orientation and category. Following the conventions in , we adopt mAP (mean Average Precision), AOS (Average Orientation Similarity) and ASE (Average Scale Error) to measure the performance of different detectors.

RoboSense三维检测任务要求检测三类主要类别(即“车辆”、“行人”和“骑行者”)的 边界框，包括位置、大小、方向和类别。遵循 中的惯例，我们采用mAP(平均精度)、AOS(平均方向相似性)和ASE(平均尺度误差)来衡量不同检测器的性能。

There are several matching criteria to define the true positive for Average Precision (AP) metric calculation. For example, [9] adopts 3D Intersection-over-Union (IoU) to match each prediction with a ground-truth box, while [4] define a match through thresholding the 2D center distance on the Bird-Eye-View ground plane. As for RoboSense detection task, we also adopt a similar distance measure. Differently, we define the threshold as a relative Proportion of ground truth Closest Collision-point Distance (CCDP) from the ego-vehicle, rather than an absolute Center Distance (CD) adopted in [4]. We claim that the localization accuracy of near obstacles’ closest collision-point is more important in low-speed driving scenarios. Then AP is calculated as the normalized area under precision-recall curve [7]. Finally, mAP is obtained by averaging over all classes and matching thresholds :

定义真正例以计算平均精度(AP)指标有几种匹配标准。例如，[9]采用三维交并比(IoU)将每个预测与真实框匹配，而[4]通过在鸟瞰图地面平面上阈值化二维中心距离来定义匹配。对于RoboSense检测任务，我们也采用了类似的距离度量。不同的是，我们将阈值定义为自车到最近碰撞点距离(CCDP)的相对比例 ，而不是[4]中采用的绝对中心距离(CD) 。我们声称在低速驾驶场景中，近处障碍物最近碰撞点的定位精度更为重要。然后AP计算为精确率-召回率曲线下的归一化面积[7]。最后，mAP通过对所有类别 和匹配阈值 取平均得到:

In addition to AP, we also measure AOS and ASE for each matched true positive, which represent the precision of predicted yaw angle and object scale respectively. AOS (Average Orientation Similarity) is formulated as:

除了AP外，我们还为每个匹配的真正例测量AOS和ASE，分别表示预测的偏航角和物体尺度的精度。AOS(平均方向相似性)的公式为:

where indicates the recall range interpolated with 40 points. indicates the set of matched true positives at recall . And denotes the angle difference between sample and ground truth. Different from [33], we only consider true positive samples under each recall level, rather than all predicted positives.

其中 表示使用40个点插值的召回范围 。 表示在召回 时匹配的真阳性样本集。而 表示样本 与真实值之间的角度差。与[33]不同，我们只考虑每个召回级别下的真阳性样本，而不是所有预测的阳性样本。

ASE is defined as , which aims to measure the scale error through calculating the after aligning orientation and translation of predictions with ground truth.

ASE定义为 ，旨在通过计算 来测量尺度误差，该误差是在将预测的方向和平移与真实值对齐后计算的。

# 4.1.2 Multi-Object Tracking

# 4.1.2 多目标跟踪

The tracking task is designed to associate all detected boxes of movable object classes across input multi-view temporal sequences (i.e. videos or point cloud sequences). Each object is assigned a unique and consistent track ID from first appearance until complete vanishing. As for performance evaluation, we refer to , and mainly adopt sAMOTA (Scaled Average Multi-Object Tracking Accuracy), AMOTP (Average Multi-Object Tracking Precision) to measure the tracking performance.

跟踪任务旨在关联输入的多视角时间序列(即视频或点云序列)中所有检测到的可移动物体类别的 框。每个物体从首次出现到完全消失都被分配一个唯一且一致的跟踪ID 。至于性能评估，我们参考 ，主要采用sAMOTA(缩放平均多目标跟踪准确率)、AMOTP(平均多目标跟踪精度)来测量 跟踪性能。

Formally, sAMOTA is defined as the mean value of sMOTA over all recalls:

形式上，sAMOTA定义为所有召回率下的sMOTA的平均值:

where and represent the number of false positives (wrongly detection), false negatives (missing detection) and identity switches at the corresponding recall , respectively. Similarly, AMOTP is the average results of MOTP among different recalls, which can be defined as:

其中 和 分别表示在相应召回率 下的假阳性(错误检测)、假阴性(漏检)和身份切换的数量。同样，AMOTP是不同召回率下MOTP的平均结果，其定义为:

where is the number of true positives at the recall , and denotes the position error of matched track at timestamp . Besides, additional metrics such as MT (Most Tracked) and ML (Most Lost) [3] are also reported.

其中 表示在召回率 下的真阳性数量， 表示在时间戳 下匹配的轨迹 的位置误差。此外，还报告了其他指标，如MT(最多跟踪)和ML(最多丢失)[3]。

# 4.2. Prediction

# 4.2. 预测

# 4.2.1 Motion Forecasting

# 4.2.1 运动预测

Based on perception results, the motion forecasting task requires to predict agents’ future trajectories. Specifically, plausible trajectories in future timesteps for each agent are forecasted as offsets to the current agent’s position. Following the standard protocols , we adopt minADE (minimum Average Displacement Error), minFDE (minimum Final Displacement Error), MR (Miss Rate) and EPA (End-to-end Prediction Accuracy) as metrics to measure the precision of motion prediction. In order to decouple the accuracy of perception and prediction, these metrics are only caculated for matched TPs (True Positives), where the matching threshold is set to of ground truth distance of the closest collision-point from the ego-vehicle. And the miss threshold of minFDE is set to for calculating the MR metric.

基于感知结果，运动预测任务需要预测智能体的未来轨迹。具体来说，每个智能体在未来 个时间步长的 条可能轨迹被预测为当前智能体位置的偏移量。遵循标准协议 ，我们采用minADE(最小平均位移误差)、minFDE(最小最终位移误差)、MR(漏检率)和EPA(端到端预测准确率)作为衡量运动预测精度的指标。为了解耦感知和预测的准确性，这些指标仅针对匹配的TP(真阳性)计算，其中匹配阈值设置为自车最近碰撞点的真实距离的 。而minFDE的漏检阈值设置为 ，用于计算MR指标。

# 4.2.2 Occupancy Prediction

# 4.2.2 占用预测

The goal of occupancy prediction task is to estimate the state of each voxel in the space. Formally, a sequence of historical frames with surround-view camera images are served as input, where and . Besides, sensor intrinsic parameters together with extrinsic parameters for each frame are also provided. Then the ground truth labels describe the voxel states separately, including occupancy state and semantic label. Three states are considered on the RoboSense dataset, including "occupied", "free" and "unknown". And the semantic label of each voxel can be one of the 3 predefined object categories or an "unknown" class to indicate general objects. Furthermore, each voxel can be also equipped with extra attributes as outputs, such as instance IDs and motion vectors, which are left as our future work.

占据预测任务的目标是估计 空间中每个体素的状态。形式上，一系列 历史帧与 环视相机图像 作为输入，其中 和 。此外，每帧的传感器内参 和外参 也被提供。然后，真实标签分别描述体素状态，包括占据状态和语义标签。在RoboSense数据集中考虑了三种状态，包括“占据”、“空闲”和“未知”。每个体素的语义标签可以是3个预定义对象类别之一，或表示一般对象的“未知”类别。此外，每个体素还可以配备额外的属性作为输出，例如实例ID和运动向量，这些将作为我们未来的工作。

To evaluate the quality of predicted occupancy, we measure the whole-scene level voxel segmentation results using IoU metric for each class. Considering the low-speed driving scenarios, we evaluate the metric under different ranges around the ego vehicle in both 3D and BEV space. Finally, is obtained through averaging over 4 classes. Moreover, evaluation is only performed on the visible vox-els from the camera view.

为了评估预测占据的质量，我们使用IoU指标测量每个类别的全场景级别体素分割结果。考虑到低速驾驶场景，我们在自车周围的不同范围内评估该指标，包括3D和BEV空间。最后， 通过对4个类别的平均得到。此外，评估仅在相机视图中可见的体素上进行。

# 5. Experiments

# 5. 实验

# 5.1. Benchmark Setup

# 5.1. 基准设置

Our RoboSense dataset contains sequences (including annotated frames) of synchronized multi-sensor data, covering 6 main categories (including 22 different locations) of outdoor or semi-closed scenarios (i.e., S1-parks, S2-scenic spots, S3-squares, S4-campuses, S5-sidewalks and S6-streets). To protect the data privacy, we conduct a series of data desensitization measures through masking the human faces and car plates as well as road signs from all sensor data. The details of RoboSense dataset composition and partitioning are listed in the Tab. 2. The RoboSense dataset is collected under various illumination, traffic flow and weather conditions, to ensure the diversity of static background and movable obstacles, thus meeting the demand of different realistic applications.

我们的RoboSense数据集包含 个序列(包括 个标注帧)的同步多传感器数据，涵盖了6个主要类别(包括22个不同地点)的户外或半封闭场景(即S1-公园，S2-景点，S3-广场，S4-校园，S5-人行道和S6-街道)。为了保护数据隐私，我们通过屏蔽所有传感器数据中的人脸、车牌和路标来进行一系列数据脱敏措施。RoboSense数据集的组成和分区细节列在表2中。RoboSense数据集在各种光照、交通流量和天气条件下收集，以确保静态背景和可移动障碍物的多样性，从而满足不同实际应用的需求。

RoboSense dataset is divided into three parts with a ratio of and , for the purpose of training, testing and validation respectively. As for the scene partition, one of the 6 collected scenes (i.e. S-6) is assigned to the testing set exclusively, while the remaining scenes are shared among all splits. Ground truth labels of training and validation sets for corresponding task are provided, together with the synchronized multi-sensor raw data. However, the testing set only provides data. Hence algorithms can merely be submitted to our online benchmark for corresponding task evaluation of testing set.

RoboSense数据集按 和 的比例分为三部分，分别用于训练、测试和验证。至于场景分区，6个收集场景中的一个(即S-6)专门分配给测试集，而其余场景在所有分割中共享。为相应任务提供训练和验证集的真实标签，以及同步的多传感器原始数据。然而，测试集仅提供数据。因此，算法只能提交到我们的在线基准进行测试集的相应任务评估。

Table 2. The details of RoboSense dataset, including the proportion of day/night data among different scenes respectively; The distribution of training/testing/validation sets; The count of synchronized sequences/frames as well as annotated 3D boxes/trajectories for each scene.

表2. RoboSense数据集的详细信息，包括不同场景中白天/夜间数据的比例；训练/测试/验证集的分布；每个场景的同步序列/帧数以及标注的3D框/轨迹的数量。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Scene-ID | Distribution | | | Ratio of Dataset | | | Num of Sequences | Num of Frames | Num of 3D Boxes | Num of Trajectories |
| Day | Night | Scene | Train | Test | Val |
| S-1 | 56% | 44% | 20% | 50% | 40% | 10% | 1.5K | 26K | 310K | 36K |
| S-2 | 69% | 31% | 30% | 2.3K | 42K | 293K | 37K |
| S-3 | 71% | 29% | 17% | 1.2K | 22K | 284K | 64K |
| S-4 | 83% | 17% | 7% | 0.5K | 9K | 144K | 22K |
| S-5 | 70% | 30% | 20% | 1.6K | 26K | 297K | 44K |
| S-6 | 22% | 78% | 6% | 0% | 100% | 0% | 0.5K | 8K | 88K | 13K |
| Total | 65% | 35% | 100% | 46% | 44% | 10% | 7.6K | 133K | 1.4M | 216K |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 场景ID | 分布 | | | 数据集比例 | | | 序列数量 | 帧数 | 3D框数量 | 轨迹数量 |
| 白天 | 夜晚 | 场景 | 训练 | 测试 | 验证 |
| S-1 | 56% | 44% | 20% | 50% | 40% | 10% | 1.5K | 26K | 310K | 36K |
| S-2 | 69% | 31% | 30% | 2.3K | 42K | 293K | 37K |
| S-3 | 71% | 29% | 17% | 1.2K | 22K | 284K | 64K |
| S-4 | 83% | 17% | 7% | 0.5K | 9K | 144K | 22K |
| S-5 | 70% | 30% | 20% | 1.6K | 26K | 297K | 44K |
| S-6 | 22% | 78% | 6% | 0% | 100% | 0% | 0.5K | 8K | 88K | 13K |
| 总计 | 65% | 35% | 100% | 46% | 44% | 10% | 7.6K | 133K | 1.4M | 216K |

Table 3. Sensor Specifications on RoboSense.

表3. RoboSense(速腾聚创)传感器规格。

|  |  |  |
| --- | --- | --- |
| Modality | Sensor | Details |
| Camera | Camera | FOV: |
| Fisheye | FOV: |
| LiDAR | Hesai Pandar40M | 64 beams, 10Hz, 384k pps FOV: |
| Zvision ML30s | 40 beams, pps FOV: |
| Livox Horizon | 40 beams, pps FOV: |
| Ultrasonics | LRU | STP-313, |
| SRU | STP-318, 5cm-200cm, 40kHz, |
| Localization | GPS & IMU | GPS, IMU, AHRS. heading, roll/pitch, , RTK positioning, update rate |

|  |  |  |
| --- | --- | --- |
| 模态 | 传感器 | 详细信息 |
| 摄像头 | 摄像头 | 视场角: |
| 鱼眼镜头 | 视场角: |
| 激光雷达 | 禾赛 Pandar40M | 64 束，10Hz，384k 点/秒 视场角: |
| 速腾聚创 ML30s | 40 束， 点/秒 视场角: |
| Livox Horizon | 40 束， 点/秒 视场角: |
| 超声波 | LRU | STP-313, |
| SRU | STP-318, 5cm-200cm, 40kHz, |
| 定位 | GPS & IMU | GPS, IMU, AHRS. 航向, 横滚/俯仰, , RTK 定位, 更新率 |

# 5.2. Sensor Specifications

# 5.2. 传感器规格

The detailed specifications of all devices are shown in Tab. 3. To cover the areas from near to farther areas, we select Cameras with different focal lengths and Field of View (FOV). Besides, 5 LiDAR sensors are installed in our data collection robot, where the top Hesai Pandar40M is served as autolabeller to provide initial annotations for the splicing points of other LiDARs. 11 Ultrasonics sensors are also installed for freespace detection to ensure safety. All devices are synchronized in time via Network Time Protocol (NTP) before data collection, we utilize a time interval of as the global timestamp, and match the frame from each device with the nearest timestamp adjacent to the global timestamp. This process ultimately yields synchronized multi-sensor data at a frame rate of 10 FPS.

所有设备的详细规格如表3所示。为了覆盖从近到远的区域，我们选择了具有不同焦距和视场角(FOV)的摄像头。此外，我们的数据采集机器人上安装了5个LiDAR传感器，其中顶部的Hesai Pandar40M作为自动标注器，为其他LiDAR的拼接点提供初始标注。还安装了11个超声波传感器用于自由空间检测，以确保安全。所有设备在数据采集前通过网络时间协议(NTP)进行时间同步，我们使用 作为全局时间戳，并将每个设备的帧与最接近全局时间戳的时间戳进行匹配。这一过程最终以10帧每秒的帧率生成同步的多传感器数据。

# 5.3. Implementation Details

# 5.3. 实现细节

For LiDAR detection task, we set the point range to , with a fixed voxel size of and for pillar-based and voxel-based methods respectively. For Image detection tasks, we use ResNet18 [11] as backbone network and the input image is resized to . For practical usages, we report performance using our proposed Closest-Collision Distance Proportion (CCDP) as matching criterion. Comparisons of different matching functions on average precision are shown in Fig. 4. As expected, when using Center Distance (CD) or IOU, objects without distance differentiation can not reflect the model capability of locating closest collision points of nearby obstacles, which is more challenging and essential for low-speed driving scenarios.

对于LiDAR检测任务，我们将点范围设置为 ，并分别为基于柱状和基于体素的方法设置固定的体素大小为 和 。对于图像检测任务，我们使用ResNet18 [11]作为骨干网络，并将输入图像调整为 。为了实际应用，我们使用我们提出的最近碰撞距离比例(CCDP)作为匹配标准来报告性能。不同匹配函数在平均精度上的比较如图4所示。正如预期的那样，当使用中心距离(CD)或IOU时，没有距离区分的对象无法反映模型在定位附近障碍物最近碰撞点方面的能力，这对于低速驾驶场景更具挑战性和重要性。

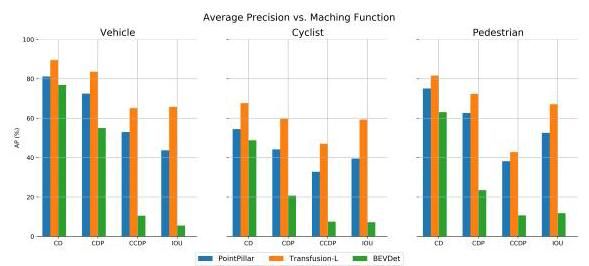


Figure 4. Average precision vs. matching function. CD: Center Distance. CDP: Center Distance Proportion. CCDP: Closest-Collision Distance Proportion. IOU: Intersection Over Union. We set IOU of Vehicle, Cyclist and Pedestrian to [0.7, 0.5, 0.5] following KITTI [9]. CD is set to following nuScenes [4] and for metrics.

图4. 平均精度与匹配函数的关系。CD:中心距离。CDP:中心距离比例。CCDP:最近碰撞距离比例。IOU:交并比。我们按照KITTI [9]将车辆、骑车人和行人的IOU设置为[0.7, 0.5, 0.5]。按照nuScenes [4]将CD设置为 ，并将 用于 指标。

# 5.4. Baselines: Perception

# 5.4. 基线:感知

# 5.4.1 LiDAR 3D Detection

# 5.4.1 LiDAR 3D检测

To demonstrate the performance of advanced 3D detectors on LiDAR-only detection track of our RoboSense benchmark, we implement several popular CNN-based methods with different fashions, including Pointpillar [17] (Pillar-based), SECOND [41] (Voxel-based), and PV-RCNN [31] (Two-stage Point-Voxel based). Besides, Transformer-based method such as Transfusion-L [2] is also implemented for architecture comparison. Pointpillar as the most efficient method above is adopted as our baseline for LiDAR 3D detection task.

为了展示先进3D检测器在我们RoboSense基准测试中仅使用LiDAR的检测轨道上的性能，我们实现了几种流行的基于CNN的方法，包括Pointpillar [17](基于柱状)、SECOND [41](基于体素)和PV-RCNN [31](基于两阶段点体素)。此外，还实现了基于Transformer的方法，如Transfusion-L [2]，用于架构比较。Pointpillar作为上述最有效的方法，被采用为我们LiDAR 3D检测任务的基线。

# 5.4.2 Multi-View 3D Detection

# 5.4.2 多视图3D检测

Current works of multi-view 3D detection can be divided into two mainstreams, namely LSS [28] based and Transformer based. To examine the effectiveness of image-only multi-view 3D detection models, we select the widely-used method BEVDet [14] as our LSS-based baseline on image 3D detection track of RoboSense, and re-implement several extended versions such as BEVDet4D [13] which takes advantage of history temporal clues, and BEVDepth [18] which adopts an additional branch for depth prediction under point supervision. Besides, BEVFormer [19] as a Transformer-based representative work is also included.

当前的多视图3D检测工作可以分为两大主流，即基于LSS [28]和基于Transformer的方法。为了检验仅使用图像的多视图3D检测模型的有效性，我们选择广泛使用的方法BEVDet [14]作为我们在RoboSense图像3D检测轨道上的基于LSS的基线，并重新实现了几个扩展版本，如BEVDet4D [13]，它利用了历史时间线索，以及BEVDepth [18]，它在点监督下采用了一个额外的分支进行深度预测。此外，还包括了基于Transformer的代表性工作BEVFormer [19]。

Table 4. 3D Detection results on validation sets of RoboSense using Center-Point (CP) distance and Closest Collision-Point (CCP) distance as matching criteria respectively where the relative proportion is set to (LiDAR) and 10% (Image).

表4. 在RoboSense验证集上使用中心点(CP)距离和最近碰撞点(CCP)距离作为匹配标准的3D检测结果，其中相对比例 设置为 (LiDAR)和10%(图像)。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Task |  | Vehicle@p=5%/10% | | | Cyclist@p=5%/10% | | | Pedestrian@ | | |
| 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ |
| LiDAR 3D Detection | PointPillar [17] | 72.5/53.0 | 73.5/61.1 | 20.6/16.1 | 44.2/32.8 | 45.4/38.3 | 64.2/54.3 | 62.7/38.2 | 45.3/34.1 | 38.3/27.2 |
| SECOND [41] | 78.8/63.1 | 80.2/69.4 | 19.8/15.7 | 53.8/43.5 | 57.2/49.9 | 67.7/55.7 | 70.8/47.2 | 54.6/43.2 | 40.1/29.3 |
| PVRCNN [31] | 74.6/57.4 | 77.4167.7 |  | 53.6/41.4 | 55.7/50.1 | 62.5/61.9 | 66.4/39.1 | 50.1/37.0 | 40.4/25.5 |
| Transfusion-L [2] |  |  | 19.7/16.0 |  |  | 82.1/72.9 | 72.3/42.8 |  | 45.1/37.4 |
| Multi-view 3D Detection | BEVDet [14] | 76.2/30.2 | 40.4/25.9 | 17.3/11.2 | 42.3/25.7 | 36.1/30.2 | 56.5/42.1 | 47.4/28.5 | 48.6/36.5 | 30.2/18.8 |
| BEVDet4D [13] | 77.2/31.1 | 41.1/26.4 | 16.8/10.8 | 42.0/24.8 | 33.9/27.7 | 55.3/41.2 | 48.1/29.3 | 46.6/37.6 |  |
| BEVDepth [18] | 77.8/31.3 | 40.9/26.3 | 16.7/10.7 | 43.3/27.0 | 34.9/30.2 |  | 50.1/31.3 | 46.7/37.9 | 28.0/21.4 |
| BEVFormer [19] | 78.2/32.0 | 41.6/26.7 | 16.5/10.6 | 44.1/27.6 | 34.9/30.5 | 51.3/44.3 | 50.2/32.3 | 46.3/38.0 |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 任务 |  | 车辆@p=5%/10% | | | 骑行者@p=5%/10% | | | 行人@ | | |
| 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ |
| LiDAR 3D检测 | PointPillar [17] | 72.5/53.0 | 73.5/61.1 | 20.6/16.1 | 44.2/32.8 | 45.4/38.3 | 64.2/54.3 | 62.7/38.2 | 45.3/34.1 | 38.3/27.2 |
| SECOND [41] | 78.8/63.1 | 80.2/69.4 | 19.8/15.7 | 53.8/43.5 | 57.2/49.9 | 67.7/55.7 | 70.8/47.2 | 54.6/43.2 | 40.1/29.3 |
| PVRCNN [31] | 74.6/57.4 | 77.4167.7 |  | 53.6/41.4 | 55.7/50.1 | 62.5/61.9 | 66.4/39.1 | 50.1/37.0 | 40.4/25.5 |
| Transfusion-L [2] |  |  | 19.7/16.0 |  |  | 82.1/72.9 | 72.3/42.8 |  | 45.1/37.4 |
| 多视角3D检测 | BEVDet [14] | 76.2/30.2 | 40.4/25.9 | 17.3/11.2 | 42.3/25.7 | 36.1/30.2 | 56.5/42.1 | 47.4/28.5 | 48.6/36.5 | 30.2/18.8 |
| BEVDet4D [13] | 77.2/31.1 | 41.1/26.4 | 16.8/10.8 | 42.0/24.8 | 33.9/27.7 | 55.3/41.2 | 48.1/29.3 | 46.6/37.6 |  |
| BEVDepth [18] | 77.8/31.3 | 40.9/26.3 | 16.7/10.7 | 43.3/27.0 | 34.9/30.2 |  | 50.1/31.3 | 46.7/37.9 | 28.0/21.4 |
| BEVFormer [19] | 78.2/32.0 | 41.6/26.7 | 16.5/10.6 | 44.1/27.6 | 34.9/30.5 | 51.3/44.3 | 50.2/32.3 | 46.3/38.0 |  |

Table 5. Study of different sensor layouts for perception tasks (3D detection and MOT) on validation sets of RoboSense under different ranges (m). AB3DMOT [39] is adopted as 3D MOT baseline. C: Camera, F: Fisheye, L: LiDAR, V: View

表5. 不同传感器布局在RoboSense验证集上用于感知任务(3D检测和多目标跟踪)的研究，范围(米)不同。采用AB3DMOT [39]作为3D多目标跟踪基线。C:相机，F:鱼眼镜头，L:激光雷达，V:视角

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Task | Detector | Layouts | Detection | | | | Tracking | | | |
|  | Range(m) | | | sAMOTA↑ | AMOTP↑ |  |  |
|  | [5, 10] | [10, 30] |
| Multi-view 3D Perception | BEVDepth [18] | 4C | 3D AP | 54.9/16.0 | 60.1/18.3 | 53.7/33.1 | 44.03 | 29.95 | 20.23 | 54.01 |
| AOS | 44.8/19.7 | 37.0/18.8 | 34.5/26.9 |
| 4F | 3D AP | 61.1/16.9 | 70.6/19.9 | 50.8/29.0 | 39.56 | 27.10 | 18.02 | 61.74 |
| AOS | 58.7/27.5 | 41.3/23.5 | 36.1/27.4 |
| 4C + 4F | 3D AP |  | 75.2/22.9 | 64.2/38.6 | 51.16 | 35.68 | 25.21 | 48.07 |
| AOS | 53.9/24.4 | 43.1/22.5 | 39.6/30.9 |
| LiDAR 3D Perception | PointPillar [17] | 4L | 3D AP | 59.2/19.3 | 73.1/42.0 | 71.0/65.4 | 44.77 | 33.65 | 25.04 | 54.08 |
| AOS | 46.5/19.2 | 67.2/47.5 | 69.0/65.7 |
| Multi-modal 3D Perception | BEVDepth [18] + Pointpillar [17] |  | 3D AP |  | 61.3/54.6 | 54.4/52.6 | 43.32 | 43.18 | 34.74 | 40.82 |
| AOS |  | 78.7/75.0 |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 任务 | 检测器 | 布局 | 检测 | | | | 跟踪 | | | |
|  | 范围(米) | | | sAMOTA↑ | AMOTP↑ |  |  |
|  | [5, 10] | [10, 30] |
| 多视角3D感知 | BEV深度 [18] | 4C | 3D AP | 54.9/16.0 | 60.1/18.3 | 53.7/33.1 | 44.03 | 29.95 | 20.23 | 54.01 |
| AOS | 44.8/19.7 | 37.0/18.8 | 34.5/26.9 |
| 4F | 3D AP | 61.1/16.9 | 70.6/19.9 | 50.8/29.0 | 39.56 | 27.10 | 18.02 | 61.74 |
| AOS | 58.7/27.5 | 41.3/23.5 | 36.1/27.4 |
| 4C + 4F | 3D AP |  | 75.2/22.9 | 64.2/38.6 | 51.16 | 35.68 | 25.21 | 48.07 |
| AOS | 53.9/24.4 | 43.1/22.5 | 39.6/30.9 |
| 激光雷达3D感知 | 点柱 [17] | 4L | 3D AP | 59.2/19.3 | 73.1/42.0 | 71.0/65.4 | 44.77 | 33.65 | 25.04 | 54.08 |
| AOS | 46.5/19.2 | 67.2/47.5 | 69.0/65.7 |
| 多模态3D感知 | BEV深度 [18] + 点柱 [17] |  | 3D AP |  | 61.3/54.6 | 54.4/52.6 | 43.32 | 43.18 | 34.74 | 40.82 |
| AOS |  | 78.7/75.0 |  |

# 5.4.3 Multiple Object Tracking

# 5.4.3 多目标跟踪

We follow the "Tracking-by-Detection" paradigm using 3D detection results from Camera or LiDAR data as input respectively, and present several baselines for multiple 3D object tracking task. Specifically, 3D boxes detected from surround-view images by BEVDepth [18] and splicing pointclouds by Pointpillar [17] are provided separately. And the tracking approach AB3DMOT described in [39] is picked to serve as the baseline of multiple object tracker in the 3D space. Then the same objects across different sensors are associated with unique track IDs to form global trajectories in the past.

我们遵循“检测跟踪”范式，分别使用相机或LiDAR数据的3D检测结果作为输入，并提出了几种多3D目标跟踪任务的基线方法。具体来说，分别提供了通过BEVDepth [18]从环视图像中检测到的3D框和通过Pointpillar [17]拼接的点云。并选择[39]中描述的AB3DMOT跟踪方法作为3D空间中多目标跟踪器的基线。然后，将不同传感器中的相同目标与唯一的轨迹ID关联，以形成过去的时间轨迹。

# 5.5. Baselines: Prediction

# 5.5. 基线:预测

# 5.5.1 Motion Prediction

# 5.5.1 运动预测

Traditional motion prediction methods utilize perception ground truth (i.e., history trajectories of agents and HDmap) as input, which lacks of uncertainty modeling in practical applications. In this paper, we implement several vision-based end-to-end methods for joint perception and motion prediction on RoboSense benchmark, including ViP3D [10] and PnPNet [20]. For comparisons, we also report the motion prediction results of assuming agents surrounding the ego-vehicle with constant positions or velocities respectively, thus to reflect the diversity and difficulty of our dataset on prediction task.

传统的运动预测方法利用感知真值(即代理的历史轨迹和高清地图)作为输入，这在实际应用中缺乏不确定性建模。在本文中，我们在RoboSense基准上实现了几种基于视觉的端到端方法，用于联合感知和运动预测，包括ViP3D [10]和PnPNet [20]。为了进行比较，我们还报告了假设代理分别以恒定位置或速度围绕自车运动的运动预测结果，从而反映我们数据集在预测任务上的多样性和难度。

# 5.5.2 Occupancy Prediction

# 5.5.2 占据预测

We extend a BEV 3D detection model - BEVDepth [18] to the 3D occupancy prediction task, which is then adopted as our baseline for the visual occupancy prediction task. Concretely, we replace the original detection decoders with the occupancy reconstruction layers while maintaining the BEV feature encoders. ResNet18 [11] pretrained on FCOS3D [38] is employed as image backbone for visual feature extraction.

我们将BEV 3D检测模型——BEVDepth [18]扩展到3D占据预测任务，并将其作为视觉占据预测任务的基线。具体来说，我们替换了原始的检测解码器，同时保留了BEV特征编码器。在FCOS3D [38]上预训练的ResNet18 [11]被用作图像骨干网络，用于视觉特征提取。

# 5.6. Results and Analysis

# 5.6. 结果与分析

# 5.6.1 Perception Results

# 5.6.1 感知结果

3D Object Detection. The 3D detection results based on multi-view images and splicing point clouds are shown in Tab. 4. As for LiDAR 3D detection, Transfusion-L [2] achieves the leading performance owing to the advanced transformer architecture. In terms of multi-view 3D detection, BEVDet4D [13] and BEVDepth [18] obtain significant improvement than BEVDet [14] through involving temporal clues and adopting an additional depth branch respectively. Besides, BEVFormer [19] also achieves competitive results by introducing a query-based attention mechanism. Generally, LiDAR-based 3D detector can generate high-quality detection results than vision-based methods. However, vision-based methods are capable of detecting various ranges of objects with more sensors (Fisheye or Camera). Note that two different matching criteria are both considered for TP calculation, namely Center-Point (CP) distance and Closest Collision-Point (CCP) distance. It can be observed that the CCP localization performance is obviously lower than the CP localization (i.e. 18.5% 3D AP drop of Transfusion-L for Vehicle class and 29.5% 3D AP drop for Pedestrian class. For navigation safety, the CCP localization is more important for near-field egocentric perception in crowded social scenarios.

3D目标检测。基于多视图图像和拼接点云的3D检测结果如表4所示。对于LiDAR 3D检测，Transfusion-L [2]由于采用了先进的Transformer架构，取得了领先的性能。在多视图3D检测方面，BEVDet4D [13]和BEVDepth [18]通过引入时间线索和采用额外的深度分支，分别比BEVDet [14]取得了显著提升。此外，BEVFormer [19]通过引入基于查询的注意力机制，也取得了有竞争力的结果。总体而言，基于LiDAR的3D检测器比基于视觉的方法能生成更高质量的检测结果。然而，基于视觉的方法能够通过更多传感器(鱼眼或相机)检测到各种范围的目标。需要注意的是，TP计算时考虑了两种不同的匹配标准，即中心点(CP)距离和最近碰撞点(CCP)距离。可以观察到，CCP定位性能明显低于CP定位(即Transfusion-L在车辆类上的3D AP下降了18.5%，在行人类上下降了29.5%)。对于导航安全而言，CCP定位在拥挤的社会场景中的近场自车感知更为重要。

Performance with Different Sensor-layouts. To evaluate the performance of different sensor layouts under various ranges, we conduct extensive comparisons as shown in Tab. 5. As for visual perception, 4C layout achieves better AP than layout in farther areas (i.e., ), while layout is good at detecting near-field targets within . Through combining these two layouts, better performance can be achieved across different ranges. LiDAR 3D detector exhibits an obvious advantage over visual detectors especially in CCP and farther object localization, while the performance of near-field objects within is inferior (19.3% vs. ). Moreover, we implement multi-modal 3D perception (8V+4L) through late-fusion strategy. Specifically, 3D detection results from multi-view 3D detector and LiDAR 3D detector are adopted for post-processing. And we can observe that the CCP-based 3D AP of objects within is remarkably boosted from 20.5% to 36.9%. And the AOS metric is also increased consistently.

不同传感器布局的性能。为了评估不同传感器布局在不同范围内的性能，我们进行了广泛的比较，如表5所示。对于视觉感知，4C布局在较远区域(即 )的AP优于 布局，而 布局在 范围内的近场目标检测上表现更好。通过结合这两种布局，可以在不同范围内实现更好的性能。LiDAR 3D检测器在CCP和较远目标定位上表现出明显优势，尤其是在 范围内的近场目标性能较差(19.3% vs. )。此外，我们通过后融合策略实现了多模态3D感知(8V+4L)。具体来说，采用了多视图3D检测器和LiDAR 3D检测器的3D检测结果进行后处理。可以观察到， 范围内目标的基于CCP的3D AP从20.5%显著提升至36.9%。AOS指标也一致提升。

Multiple Object Tracking. Regarding to the MOT task in Tab. 5, AB3DMOT [39] is adopted as baseline tracker in 3D space, which mitigates the impact of object occlusions existing in image, especially for crowded scenarios. Through introducing more sensors , vision-based methods can also achieve competitive tracking performance with LiDAR-based methods, even better in sAMOTA metric (51.16 vs. 44.77). With the multi-modal input, AMOTP, MT and ML performance can be further improved as expected. However, although equipped with multi-modal and multi-sensor data as input, the perception performance is still inferior especially in near-field (i.e. 36.9% CCP-based 3D AP within ), revealing the deficiencies of current perception methods in handling the obstacles in near ranges. The main reason may be the frequent truncation and occlusion caused by a large view occupation of near

多目标跟踪。关于表5中的MOT任务，AB3DMOT [39]被采用为3D空间中的基线跟踪器，它减轻了 图像中存在的目标遮挡的影响，特别是在拥挤场景中。通过引入更多传感器 ，基于视觉的方法也可以实现与基于LiDAR的方法相媲美的跟踪性能，甚至在sAMOTA指标上表现更好(51.16 vs. 44.77)。通过多模态输入，AMOTP、MT和ML性能可以如预期进一步改善。然而，尽管配备了多模态和多传感器数据作为输入，感知性能仍然较差，特别是在近场(即在 内的基于CCP的3D AP为36.9%)，揭示了当前感知方法在处理近距离障碍物方面的不足。主要原因可能是近场大视角占用导致的频繁截断和遮挡。

Table 6. Motion forecasting results on validation sets of Ro-boSense. and indicate utilizing GroundTruth 3D boxes and detection results from PointPillar [17] as input respectively with constant positions or velocities for comparisons.

表6. Ro-boSense验证集上的运动预测结果。 和 分别表示使用GroundTruth 3D框和PointPillar [17]的检测结果作为输入，并采用恒定位置或速度进行比较。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | minADE (m) | minFDE (m) |  | EPA↑ |
| Constant Pos.\* | 2.42 | 3.01 | 0.319 | 0.680 |
| Constant Vel.\* | 1.59 | 3.54 | 0.219 | 0.780 |
| Constant Pos. | 1.52 | 1.95 | 0.267 | 0.243 |
| Random Vel. | 2.56 | 3.85 | 0.872 | 0.029 |
| ViP3D [10] | 1.31 | 1.55 | 0.196 | 0.283 |
| PnPNet [20] | 0.89 | 1.12 | 0.172 | 0.313 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 最小平均位移误差 (m) | 最小最终位移误差 (m) |  | EPA↑ |
| 恒定位置\* | 2.42 | 3.01 | 0.319 | 0.680 |
| 恒定速度\* | 1.59 | 3.54 | 0.219 | 0.780 |
| 恒定位置 | 1.52 | 1.95 | 0.267 | 0.243 |
| 随机速度 | 2.56 | 3.85 | 0.872 | 0.029 |
| ViP3D [10] | 1.31 | 1.55 | 0.196 | 0.283 |
| PnPNet [20] | 0.89 | 1.12 | 0.172 | 0.313 |

Table 7. Occupancy prediction results on validation sets of Ro-boSense using 4F sensors as input. "mIoU-3D" and "mIoU-BEV" indicate the standard mIoU metric calculated in 3D space and BEV respectively without considering the ground voxels.

表7. 使用4F传感器作为输入的Ro-boSense验证集上的占用预测结果。"mIoU-3D"和"mIoU-BEV"分别表示在3D空间和BEV(鸟瞰图)中计算的标准mIoU指标，不考虑地面体素。

|  |  |  |
| --- | --- | --- |
| Range(m) | mIoU-3D↑ | mIoU-BEV↑ |
|  | 24.6 | 29.7 |
|  | 39.6 | 48.2 |
|  | 30.7 | 36.7 |
|  | 16.1 | 19.7 |

|  |  |  |
| --- | --- | --- |
| 范围(米) | 三维平均交并比↑ | 鸟瞰图平均交并比↑ |
|  | 24.6 | 29.7 |
|  | 39.6 | 48.2 |
|  | 30.7 | 36.7 |
|  | 16.1 | 19.7 |

obstacles, which showcases the great challenge and importance of our proposed benchmark for the development of egocentric perceptual frameworks related to navigation in crowded and unstructured environments.

障碍物，这展示了我们提出的基准在开发与拥挤和非结构化环境中导航相关的自我中心感知框架方面的巨大挑战和重要性。

# 5.6.2 Prediction Results

# 5.6.2 预测结果

Motion forecasting of surrounding agents as well as occupancy state descriptions around the ego-vehicle are two crucial prediction tasks in the research field of autonomous driving, which have been extensively explored in urban and highway scenarios for autonomous cars.

周围智能体的运动预测以及自我车辆周围的占用状态描述是自动驾驶研究领域中的两个关键预测任务，这些任务在城市和高速公路场景中已经得到了广泛探索。

Motion Prediction. As shown in Tab. 6, either visual end-to-end methods [10] or LiDAR-based end-to-end methods [20] are all supported for validation on our RoboSense. PnPNet [20] with LiDAR points as input can produce less prediction errors and better EPA than ViP3D [10], both of which remarkably outperform two baseline settings of modeling agents with constant positions or velocities.

运动预测。如表6所示，无论是视觉端到端方法[10]还是基于LiDAR的端到端方法[20]，都可以在我们的RoboSense上进行验证。以LiDAR点作为输入的PnPNet[20]比ViP3D[10]产生更少的预测误差和更好的EPA，两者都显著优于以恒定位置或速度建模智能体的两个基线设置。

Occupancy Prediction. As shown in Tab. 7, we use 4F sensor data as input and report the performance of mIOU metric in both 3D and BEV space under various ranges respectively. Note that the metric is calculated without considering states of the ground voxels, leading to lower performance in either 3D or BEV space. As expected, the performance evaluated within is better than farther areas.

占用预测。如表7所示，我们使用4F传感器数据作为输入，并分别报告了在不同范围内的3D和BEV空间中的mIOU指标性能。请注意，该指标的计算不考虑地面体素的状态，导致在3D或BEV空间中的性能较低。正如预期的那样，在 范围内评估的性能优于更远的区域。

# 6. Conclusion

# 6. 结论

To foster the research of egocentric perceptual framework tailored to various types of autonomous agents navigating in crowded and unstructured environments, RoboSense, a real-world and multi-modal dataset is collected in complex social scenarios with varying and uncontrolled environmental conditions and dynamical elements. It consists of scenes manually selected from different locations, with 1.4M 3D Boxes and 216K trajectories annotated in total on synchronous frames. Besides, occupancy descriptions are also provided to facilitate the surrounding context comprehension. In the future works, more tasks and associated benchmarks, such as motion planning, will be expanded for end-to-end autonomous navigating application, and explore the additional benefits that joint optimization can bring to the modular training.

为了促进针对在拥挤和非结构化环境中导航的各种类型自主智能体的自我中心感知框架的研究，RoboSense，一个真实世界和多模态的数据集，在复杂的社会场景中收集，这些场景具有变化和不受控制的环境条件和动态元素。它由从不同地点手动选择的 个场景组成，总共有1.4M个3D框和216K条轨迹在 个同步帧上进行了标注。此外，还提供了占用描述以促进周围环境的理解。在未来的工作中，将扩展更多任务和相关基准，如运动规划，用于端到端的自主导航应用，并探索联合优化对模块化训练的额外好处。

# References

# 参考文献

[1] Ben Agro, Quinlan Sykora, Sergio Casas, and Raquel Urta-sun. Implicit occupancy flow fields for perception and prediction in self-driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1379-1388, 2023.

Ben Agro, Quinlan Sykora, Sergio Casas, 和 Raquel Urta-sun。自动驾驶中的隐式占用流场用于感知和预测。在IEEE/CVF计算机视觉与模式识别会议论文集，第1379-1388页，2023年。

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for object detection with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1090-1099, 2022.6,7,3

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, 和 Chiew-Lan Tai。Transfusion:用于 目标检测的鲁棒LiDAR-相机融合与变压器。在IEEE/CVF计算机视觉与模式识别会议论文集，第1090-1099页，2022年。6,7,3

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing, 2008:1- 10, 2008. 5

Keni Bernardin 和 Rainer Stiefelhagen。评估多目标跟踪性能:CLEAR MOT指标。EURASIP图像与视频处理期刊，2008年:1-10，2008年。5

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Gi-ancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11621-11631, 2020. 1, 2, 3, 4, 5, 6

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Gi-ancarlo Baldan, 和 Oscar Beijbom。nuscenes:用于自动驾驶的多模态数据集。在IEEE/CVF计算机视觉与模式识别会议论文集，第11621-11631页，2020年。1, 2, 3, 4, 5, 6

[5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jag-jeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8748-8757, 2019. 2, 3

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jag-jeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, 等。Argoverse:具有丰富地图的3D跟踪和预测。在IEEE/CVF计算机视觉与模式识别会议论文集，第8748-8757页，2019年。2, 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213-3223, 2016. 2, 3

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, 和 Bernt Schiele。用于语义城市场景理解的Cityscapes数据集。在IEEE计算机视觉与模式识别会议论文集，第3213-3223页，2016年。2, 3

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88:303-338, 2010. 4

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, 和 Andrew Zisserman。Pascal视觉对象类别(VOC)挑战。国际计算机视觉期刊，88:303-338，2010年。4

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition, pages 3354-3361. IEEE, 2012. 1, 2, 3

Andreas Geiger, Philip Lenz, 和 Raquel Urtasun. 我们准备好迎接自动驾驶了吗？KITTI视觉基准套件。在2012年IEEE计算机视觉与模式识别会议上，页码3354-3361。IEEE, 2012. 1, 2, 3

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11):1231-1237, 2013.4,5.6

Andreas Geiger, Philip Lenz, Christoph Stiller, 和 Raquel Urtasun. 视觉与机器人学的结合:KITTI数据集。国际机器人研究杂志，32(11):1231-1237, 2013.4,5.6

[10] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5496-5506, 2023. 5, 7, 8

Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, 和 Hang Zhao. VIP3D:通过3D代理查询实现端到端的视觉轨迹预测。在IEEE/CVF计算机视觉与模式识别会议论文集，页码5496-5506, 2023. 5, 7, 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778, 2016. 6, 7

Kaiming He, Xiangyu Zhang, Shaoqing Ren, 和 Jian Sun. 深度残差学习用于图像识别。在IEEE计算机视觉与模式识别会议论文集，页码770-778, 2016. 6, 7

[12] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17853-17862, 2023.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, 等. 面向规划的自动驾驶。在IEEE/CVF计算机视觉与模式识别会议论文集，页码17853-17862, 2023.

[13] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054, 2022. 7, 8, 3

Junjie Huang 和 Guan Huang. BEVDet4D:在多摄像头3D物体检测中利用时间线索。arXiv预印本arXiv:2203.17054, 2022. 7, 8, 3

[14] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790, 2021. 6, 7, 8, 3

Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, 和 Dalong Du. BEVDet:鸟瞰图中的高性能多摄像头3D物体检测。arXiv预印本arXiv:2112.11790, 2021. 6, 7, 8, 3

[15] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 954-960, 2018. 1, 2, 3

Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, 和 Ruigang Yang. 用于自动驾驶的ApolloScape数据集。在IEEE计算机视觉与模式识别研讨会论文集，页码954-960, 2018. 1, 2, 3

[16] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platin-sky, W. Jiang, and V. Shet. Lyft level 5 av dataset 2019. https://level5.lyft.com/dataset/, 2019. 2, 3

R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platin-sky, W. Jiang, 和 V. Shet. Lyft Level 5 AV数据集2019. https://level5.lyft.com/dataset/, 2019. 2, 3

[17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12697-12705, 2019. 6, 7, 8, 2, 3

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, 和 Oscar Beijbom. PointPillars:从点云中进行物体检测的快速编码器。在IEEE/CVF计算机视觉与模式识别会议论文集，页码12697-12705, 2019. 6, 7, 8, 2, 3

[18] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 1477-1485, 2023. 7, 8, 3

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, 和 Zeming Li. BEVDepth:为多视图3D物体检测获取可靠的深度。在AAAI人工智能会议论文集，页码1477-1485, 2023. 7, 8, 3

[19] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chong-hao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In European conference on computer vision, pages 1-18. Springer, 2022. 7, 8, 3

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chong-hao Sima, Tong Lu, Yu Qiao, 和 Jifeng Dai. BEVFormer:通过时空变换器从多摄像头图像中学习鸟瞰图表示。在欧洲计算机视觉会议上，页码1-18. Springer, 2022. 7, 8, 3

[20] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11553-11562, 2020. 5, 7, 8

梁明, 杨斌, 曾文远, 陈云, 胡锐, Sergio Casas, 和 Raquel Urtasun. Pnpnet: 端到端的感知与预测，带循环跟踪. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码11553-11562, 2020. 5, 7, 8

[21] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 3569-3577, 2018. 5

罗文杰, 杨斌, 和 Raquel Urtasun. 快而狂: 实时端到端3D检测、跟踪与运动预测，使用单一卷积网络. 在IEEE计算机视觉与模式识别会议论文集, 页码3569-3577, 2018. 5

[22] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. Artificial intelligence, 293:103448, 2021. 5

罗文瀚, 邢俊良, Anton Milan, 张小琴, 刘伟, 和 Tae-Kyun Kim. 多目标跟踪: 文献综述. 人工智能, 293:103448, 2021. 5

[23] Cong Ma, Lei Qiao, Chengkai Zhu, Kai Liu, Zelong Kong, Qing Li, Xueqi Zhou, Yuheng Kan, and Wei Wu. Holovic: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22129-22138, 2024.

马聪, 乔磊, 朱成凯, 刘凯, 孔泽龙, 李青, 周雪琪, 阚宇恒, 和吴伟. Holovic: 大规模多传感器全息交叉与车路协同数据集与基准. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码22129-22138, 2024.

[24] Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Li-udmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. arXiv preprint arXiv:2107.07455, 2021. 2

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Li-udmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, 等. Shifts: 跨多个大规模任务的真实分布偏移数据集. arXiv预印本 arXiv:2107.07455, 2021. 2

[25] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In 2019 International Conference on Robotics and Automation (ICRA), pages 9552-9557. IEEE, 2019. 2, 3

Abhishek Patil, Srikanth Malla, 杭海明, 和 陈一婷. H3D数据集: 用于拥挤城市场景中全环绕3D多目标检测与跟踪. 在2019年国际机器人与自动化会议(ICRA), 页码9552-9557. IEEE, 2019. 2, 3

[26] Neehar Peri, Jonathon Luiten, Mengtian Li, Aljoša Ošep, Laura Leal-Taixé, and Deva Ramanan. Forecasting from lidar via future object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17202-17211, 2022. 5

Neehar Peri, Jonathon Luiten, 李梦天, Aljoša Ošep, Laura Leal-Taixé, 和 Deva Ramanan. 通过未来目标检测从激光雷达进行预测. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码17202-17211, 2022. 5

[27] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A\* 3d dataset: Towards autonomous driving in challenging environments. In 2020 IEEE International conference on Robotics and Automation (ICRA), pages 2267-2273. IEEE, 2020. 1

Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, 詹慧静, 庞春浩, 陈宇达, Armin Mustafa, Vijay Chandrasekhar, 和林杰. A\* 3D数据集: 面向挑战性环境中的自动驾驶. 在2020年IEEE国际机器人与自动化会议(ICRA), 页码2267-2273. IEEE, 2020. 1

[28] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV 16, pages 194-210. Springer, 2020. 6

Jonah Philion 和 Sanja Fidler. 提升、展开、射击: 通过隐式反投影到3D从任意相机装置编码图像. 在计算机视觉-ECCV 2020: 第16届欧洲会议, 英国格拉斯哥, 2020年8月23-28, 论文集, 第XIV部分16, 页码194-210. Springer, 2020. 6

[29] Davide Scaramuzza, Agostino Martinelli, and Roland Sieg-wart. A flexible technique for accurate omnidirectional camera calibration and structure from motion. In Fourth IEEE International Conference on Computer Vision Systems (ICVS’06), pages 45-45. IEEE, 2006. 1

Davide Scaramuzza, Agostino Martinelli, 和 Roland Siegwart. 一种灵活的技术，用于精确的全向相机校准与运动结构. 在第四届IEEE国际计算机视觉系统会议(ICVS’06), 页码45-45. IEEE, 2006. 1

[30] Gregory C Sharp, Sang W Lee, and David K Wehe. Icp registration using invariant features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1):90-102, 2002. 3

Gregory C Sharp, Sang W Lee, 和 David K Wehe. 使用不变特征的ICP配准. IEEE模式分析与机器智能汇刊, 24(1):90-102, 2002. 3

[31] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages

师少帅, 郭超旭, 姜立, 王哲, 师建平, 王晓刚, 和李洪升. Pv-rcnn: 点-体素特征集抽象用于 目标检测. 在IEEE/CVF计算机视觉与模式识别会议论文集,页码

[32] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460, 2020. 5

孙培泽, 曹金坤, 蒋毅, 张如峰, 谢恩泽, 袁泽寰, 王昌虎, 罗平. Transtrack: 使用Transformer进行多目标跟踪. arXiv预印本 arXiv:2012.15460, 2020. 5

[33] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception

孙培, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, 周寅, 柴宇宁, Benjamin Caine, 等. 自动驾驶感知的可扩展性

for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2446-2454, 2020. 1, 2, 3, 4, 5

Waymo开放数据集. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码2446-2454, 2020. 1, 2, 3, 4, 5

[34] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xi-angxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15035-15044, 2024. 2

唐品, 王忠道, 王国庆, 郑继来, 任向轩, 冯柏兰, 马超. Sparseocc: 重新思考基于视觉的语义占用预测中的稀疏潜在表示. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码15035-15044, 2024. 2

[35] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. Advances in Neural Information Processing Systems, 36, 2024. 4

田晓宇, 蒋涛, 云龙飞, 毛宇成, 杨慧彤, 王悦, 王一伦, 赵航. Occ3d: 自动驾驶的大规模3D占用预测基准. 神经信息处理系统进展, 36, 2024. 4

[36] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8406- 8415, 2023. 4

童文文, 司马崇浩, 王泰, 陈立, 吴思磊, 邓汉明, 顾毅, 卢乐伟, 罗平, 林达华, 等. 场景作为占用. 在IEEE/CVF国际计算机视觉会议论文集, 页码8406-8415, 2023. 4

[37] Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xi-angxuan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. arXiv preprint arXiv:2404.15014, 2024.

王国庆, 王忠道, 唐品, 郑继来, 任向轩, 冯柏兰, 马超. Occgen: 自动驾驶的生成式多模态3D占用预测. arXiv预印本 arXiv:2404.15014, 2024.

[38] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 913-922, 2021. 7

王泰, 朱新歌, 庞江淼, 林达华. Fcos3d: 全卷积单阶段单目3D目标检测. 在IEEE/CVF国际计算机视觉会议论文集, 页码913-922, 2021. 7

[39] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for multi-object tracking and new evaluation metrics. arXiv e-prints, 2020. 7, 8

翁新硕, 王建仁, David Held, Kris Kitani. Ab3dmot: 多目标跟踪的基线和新评估指标. arXiv电子版, 2020. 7, 8

[40] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pages 3095-3101. IEEE, 2021. 1

肖鹏川, 邵振磊, Steven Hao, 张子硕, 柴晓林, 焦朱迪, 李泽松, 吴健, 孙凯, 江坤, 等. Pandaset: 自动驾驶的高级传感器套件数据集. 在2021年IEEE国际智能交通系统会议(ITSC), 页码3095-3101. IEEE, 2021. 1

[41] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. Sensors, 18(10):3337, 2018. 6, 7, 3

严岩, 毛宇星, 李波. Second: 稀疏嵌入卷积检测. 传感器, 18(10):3337, 2018. 6, 7, 3

[42] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2636-2645, 2020. 2, 3

余飞, 陈浩峰, 王鑫, 先文琪, 陈莹莹, 刘芳辰, Vashisht Madhavan, Trevor Darrell. Bdd100k: 异构多任务学习的多样化驾驶数据集. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码2636-2645, 2020. 2, 3

# CRoboSense: Large-scale Dataset and Benchmark for Egocentric Robot Perception and Navigation in Crowded and Unstructured Environments

# CRoboSense: 大规模数据集和基准，用于拥挤和非结构化环境中的自我中心机器人感知与导航

Supplementary Material

补充材料

# A. Coordinates Transformation

# A. 坐标变换

# A.1. LiDAR Ego-Vehicle

# A.1. LiDAR 自我车辆

LiDAR to Ego-Vehicle: represents a three-dimensional coordinate point in Ego-Vehicle Coordinate System. The transformation from the coordinates in the Ego-Vehicle Coordinate System to in the LiDAR Coordinate System is calculated as follows:

LiDAR到自车坐标系: 表示自车坐标系中的一个三维坐标点。从自车坐标系中的坐标 到LiDAR坐标系中的坐标 的转换计算如下:

where and represent the rotation and translation from the Ego-Vehicle Coordinate System to the LiDAR Coordinate System, respectively.

其中 和 分别表示从自车坐标系到LiDAR坐标系的旋转和平移。

Ego-Vehicle to LiDAR: The transformation from Ego-Vehicle Coordinate System to LiDAR Coordinate System is the inverse transformation of Eq.(1).

自车到LiDAR:从自车坐标系到LiDAR坐标系的转换是公式(1)的逆变换。

# A.2. LiDAR Camera

# A.2. LiDAR 相机

LiDAR to Camera: Regardless of whether it is a fisheye or a pinhole camera, the coordinate transformation formula from the LiDAR Coordinate System to the Camera Coordinate System is the same and is given as follows:

LiDAR到相机:无论是鱼眼相机还是针孔相机，从LiDAR坐标系到相机坐标系的坐标转换公式是相同的，如下所示:

where represents a three-dimensional coordinate point in the Camera Coordinate System. and represent the rotation and translation from the LiDAR Coordinate System to the Camera Coordinate System, respectively.

其中 表示相机坐标系中的一个三维坐标点。 和 分别表示从LiDAR坐标系到相机坐标系的旋转和平移。

Camera to LiDAR: The transformation from Camera Coordinate System to LiDAR Coordinate System is the inverse transformation of Eq.(2).

相机到LiDAR:从相机坐标系到LiDAR坐标系的转换是公式(2)的逆变换。

# A.3. Camera Pixel

# A.3. 相机 像素

Camera to Pixel: The projection formulas of different types of cameras are different in the RoboSense dataset, the

相机到像素:在RoboSense数据集中，不同类型相机的投影公式不同，

projection formula of a pinhole camera is as follows:

针孔相机的投影公式如下:

(3)

(u, v)is pixel coordinate, represents the camera intrinsic parameters, represents the focal lengths of the camera, and indicates the displacement of the camera’s optical center from the origin of the Pixel Coordinate System. The projection formula from camera coordinate to pixel coordinate of the fisheye camera is very different, the camera projection process refers to the projection formula of Omnidirectional Camera (OCam) in [29].

(u, v)是像素坐标， 表示相机的内参， 表示相机的焦距， 表示相机光心与像素坐标系原点的位移。鱼眼相机从相机坐标到像素坐标的投影公式与针孔相机有很大不同，相机投影过程参考[29]中的全向相机(OCam)的投影公式。

Pixel to Camera: The transformation from Pixel Coordinate System to Camera Coordinate System in a pinhole camera model requires the inverse of Eq.(3). Since this is a to transformation, it is necessary to first determine the magnitude of . The projection formula from pixel coordinate to camera coordinate of the fisheye camera refers to the projection formula of Omnidirectional Camera (OCam) in [29].

像素到相机:在针孔相机模型中，从像素坐标系到相机坐标系的转换需要公式(3)的逆变换。由于这是从 到 的转换，首先需要确定 的大小。鱼眼相机从像素坐标到相机坐标的投影公式参考[29]中的全向相机(OCam)的投影公式。

# A.4. Ego-Vehicle Global

# A.4. 自车 全局

Ego-Vehicle to Global: and represent the transformation matrices of the vehicle’s orientation and position in the Global Coordinate System, respectively. The transformation formula for converting the coordinates in the Ego-Vehicle Coordinate System to in the Global Coordinate System is as follows:

自车到全局: 和 分别表示车辆在全局坐标系中的方向和位置的变换矩阵。将自车坐标系中的坐标 转换为全局坐标系中的坐标 的转换公式如下:

Global to Ego-Vehicle: The transformation from Global Coordinate System to Ego-Vehicle Coordinate System is the inverse transformation of Eq.(4).

全局到自车:从全局坐标系到自车坐标系的转换是公式(4)的逆变换。

# B. More Details of RoboSense

# B. RoboSense的更多细节

# B.1. Annotation Statistics

# B.1. 标注统计

We present more statistics on the annotations of RoboSense as shown in Tab. 8. It can be observed that our Ro-boSense dataset contains approximately annotated objects, with vehicles and pedestrians comprising the majority, while cyclists are lesser. The distribution of objects is relatively uniform in terms of distance. Additionally, due to the smaller coverage area of Livox pointclouds (Local view) compared to Hesai pointclouds (Global view), the number of annotated objects in the Livox pointclouds is only of that in the Hesai pointclouds. In Fig. A1, we further compare the distribution of annotated objects between our Robosense dataset and nuScenes dataset. It is obvious that our Robosense dataset contains significantly more annotated objects of vehicles, pedestrians, and cyclists classes respectively, which tend to be closer to the ego robot.

我们提供了关于RoboSense标注的更多统计数据，如表8所示。可以观察到，我们的RoboSense数据集包含大约 个标注对象，其中车辆和行人占大多数，而骑行者较少。在距离方面，对象的分布相对均匀。此外，由于Livox点云(局部视图)的覆盖面积比Hesai点云(全局视图)小，Livox点云中的标注对象数量仅为Hesai点云的 。在图A1中，我们进一步比较了Robosense数据集和nuScenes数据集中标注对象的分布。显然，我们的Robosense数据集分别包含显著更多的车辆、行人和骑行者类别的标注对象，这些对象往往更接近自机器人。

Table 8. The Number and proportion of 3D Boxes from all sensors (Global Scenes) and Livox LiDAR (Local Scenes) per category under different ranges (m) respectively.

表8. 不同范围(米)下，所有传感器(全局场景)和Livox LiDAR(局部场景)每类别的3D框数量及比例。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Global/Local | Vehicle | | | Cyclist | | | Pedestrian | | | Total |
| [0 - 10] | [10 - 30] | [30 - ] | [0 - 10] | [10 - 30] | [30 - ] | [0 - 10] | [10 - 30] | [30 - ] |
| Global (Hesai LiDAR) | 165K | 402K | 343K | 23K | 38K | 15K | 187K | 163K | 51K | 1.4M |
| 910K | | | 76K | | | 401K | | |
| 65.00% | | | 5.42% | | | 28.64% | | | 100% |
| Local (Livox LiDAR) | 150K | 282K | 133K | 20K | 28K | 7K | 163K | 103K | 21K | 907K |
| 565K | | | 55K | | | 287K | | |
| 40.36% | | | 3.93% | | | 20.50% | | | 64.79% |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 全局/局部 | 车辆 | | | 骑行者 | | | 行人 | | | 总计 |
| [0 - 10] | [10 - 30] | [30 - ] | [0 - 10] | [10 - 30] | [30 - ] | [0 - 10] | [10 - 30] | [30 - ] |
| 全局(禾赛激光雷达) | 165K | 402K | 343K | 23K | 38K | 15K | 187K | 163K | 51K | 1.4M |
| 910K | | | 76K | | | 401K | | |
| 65.00% | | | 5.42% | | | 28.64% | | | 100% |
| 局部(览沃激光雷达) | 150K | 282K | 133K | 20K | 28K | 7K | 163K | 103K | 21K | 907K |
| 565K | | | 55K | | | 287K | | |
| 40.36% | | | 3.93% | | | 20.50% | | | 64.79% |

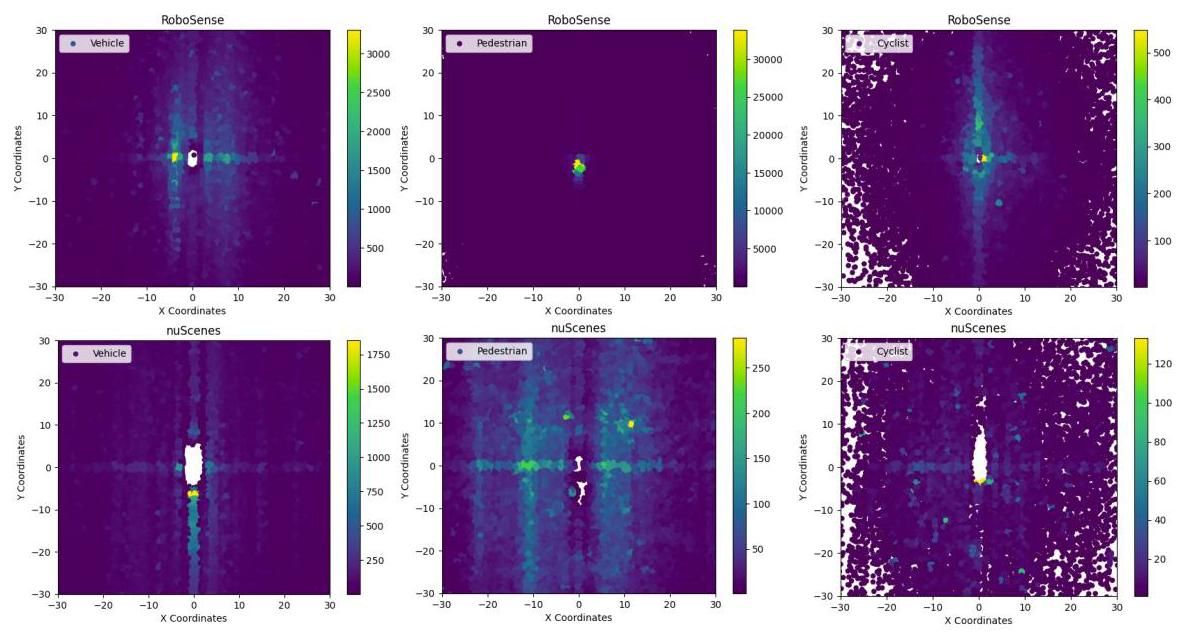


Figure A1. Comparison of annotated object distribution of different classes between RoboSense and nuScenes datasets.

图A1. RoboSense与nuScenes数据集中不同类别的标注对象分布对比。

# B.2.3D Object Label Generation

# B.2.3D 物体标签生成

To generate high-quality 3D object annotations, we design a three-stage 3D object generation pipeline for different sensors covering various ranges. First, a pre-trained LiDAR detection model (i.e., [17]) of high precision is adopted to produce 3D objects on the full view using high-quality Pandar64 points as input. Then expert annotators are required to refine the initial 3D boxes continuously throughout the whole sequences in each scene, based on splicing pointclouds which are obtained by aligning 4 vehicle-side LiDARs to the Ego-Vehicle coordinate through affine transformation. Besides, annotators need to supplement surrounding 3D boxes in a near range which are not scanned by the top Hesai LiDAR or fail to be detected owing to high occlusion and truncation. Last but not least, invalid 3D annotations should be excluded for target LiDAR coordinate and Camera coordinate respectively, where the annotated objects are not covered in the corresponding sensor data. Through multiple validation steps, highly accurate annotations can be achieved in both near and far ranges. We also release intermediate Pandar64 points for research usages.

为了生成高质量的3D物体标注，我们设计了一个三阶段的3D物体生成流程，适用于覆盖不同范围的多种传感器。首先，采用高精度的预训练LiDAR检测模型(即[17])，以高质量的Pandar64点云作为输入，在全 视图中生成3D物体。然后，专家标注者需要基于通过仿射变换将4个车载LiDAR对齐到自车坐标系获得的拼接点云，在每个场景的整个序列中持续优化初始的3D框。此外，标注者还需要补充近范围内未被顶部Hesai LiDAR扫描或由于高遮挡和截断未能检测到的周围3D框。最后但同样重要的是，应分别排除目标LiDAR坐标系和相机坐标系中的无效3D标注，这些标注对象未包含在相应的传感器数据中。通过多个验证步骤，可以在近远范围内实现高度准确的标注。我们还发布了中间Pandar64点云以供研究使用。

# B.3. Occupancy Label Preprocess

# B.3. 占据标签预处理

Occupancy label generation can be primarily divided into two parts: pointclouds densification and occupancy label determination. Unlike existing counterpart [34] which only utilizes the sparse keyframe LiDAR points, multi-frame aggregation operation is found to be indispensable for dense occupancy generation. For dynamic objects, the extracted dynamic points of neighboring frames are subsequently concatenating for each object along the corresponding trajectory respectively, thus achieving the pointclouds densification. For static scenes, coordinate transformation is performed from the ego-vehicle coordinate to the global coordinate across time using ego-pose information, and then simply aggregate all static points on the ego-vehicle coordinate of current keyframe through concatenation.

占据标签生成主要分为两部分:点云密集化和占据标签确定。与现有方法[34]仅利用稀疏关键帧LiDAR点云不同，多帧聚合操作被发现对于密集占据生成是必不可少的。对于动态物体，提取的相邻帧动态点云随后分别沿相应轨迹进行拼接，从而实现点云密集化。对于静态场景，使用自车姿态信息将自车坐标系转换为全局坐标系，然后通过拼接简单聚合当前关键帧自车坐标系上的所有静态点云。

Table 9. 3D Detection results of different modalities on validation sets of RoboSense using as matching criteria.

表9. 使用 作为匹配标准，不同模态在RoboSense验证集上的3D检测结果。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Task |  | Vehicle@IoU=0.7/0.3 | | | Cyclist@IoU=0.5/0.3 | | | Pedestrian@IoU=0.5/0.3 | | |
| 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ |
| LiDAR 3D Detection | PointPillar [17] | 43.7 | 45.5 | 13.3 | 39.5 | 39.6 | 69.2 | 52.6 | 36.6 | 34.9 |
| SECOND [41] | 55.8 | 59.8 | 17.2 | 52.3 | 53.3 | 65.9 | 61.7 | 46.9 | 37.5 |
| PVRCNN [31] | 53.5 | 57.9 | 16.9 | 53.0 | 50.7 | 55.9 | 58.9 | 43.4 | 38.4 |
| Transfusion-L [2] | 65.8 | 66.3 | 17.3 | 59.3 | 71.0 | 78.5 | 67.1 | 56.0 | 42.7 |
| Multi-view 3D Detection | BEVDet [14] | 32.1 | 21.8 | 10.4 | 19.9 | 21.2 | 36.8 | 25.9 | 29.7 | 20.3 |
| BEVDet4D [13] | 33.5 | 22.8 | 10.4 | 20.1 | 21.1 | 36.7 | 26.2 | 28.3 | 17.7 |
| BEVDepth [18] | 33.4 | 22.8 | 10.2 | 22.6 | 22.2 | 41.6 | 27.7 | 28.1 | 17.9 |
| BEVFormer [19] | 33.6 | 23.0 | 10.3 | 23.4 | 22.1 | 35.3 | 28.0 | 29.5 | 17.8 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 任务 |  | 车辆@IoU=0.7/0.3 | | | 骑行者@IoU=0.5/0.3 | | | 行人@IoU=0.5/0.3 | | |
| 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ | 3D AP↑ | AOS↑ | ASE↓ |
| LiDAR 3D检测 | PointPillar [17] | 43.7 | 45.5 | 13.3 | 39.5 | 39.6 | 69.2 | 52.6 | 36.6 | 34.9 |
| SECOND [41] | 55.8 | 59.8 | 17.2 | 52.3 | 53.3 | 65.9 | 61.7 | 46.9 | 37.5 |
| PVRCNN [31] | 53.5 | 57.9 | 16.9 | 53.0 | 50.7 | 55.9 | 58.9 | 43.4 | 38.4 |
| Transfusion-L [2] | 65.8 | 66.3 | 17.3 | 59.3 | 71.0 | 78.5 | 67.1 | 56.0 | 42.7 |
| 多视角3D检测 | BEVDet [14] | 32.1 | 21.8 | 10.4 | 19.9 | 21.2 | 36.8 | 25.9 | 29.7 | 20.3 |
| BEVDet4D [13] | 33.5 | 22.8 | 10.4 | 20.1 | 21.1 | 36.7 | 26.2 | 28.3 | 17.7 |
| BEVDepth [18] | 33.4 | 22.8 | 10.2 | 22.6 | 22.2 | 41.6 | 27.7 | 28.1 | 17.9 |
| BEVFormer [19] | 33.6 | 23.0 | 10.3 | 23.4 | 22.1 | 35.3 | 28.0 | 29.5 | 17.8 |

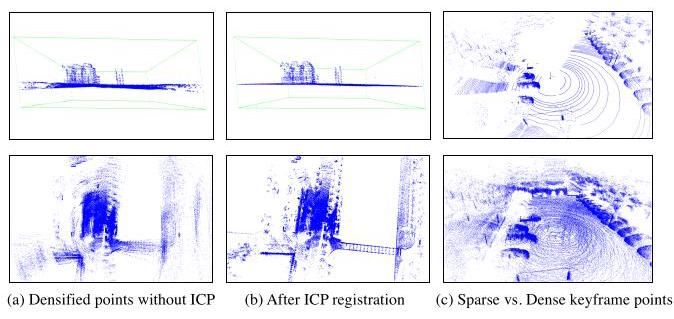


Figure A2. Illustration of ICP and points densified process.

图A2. ICP和点云密集化过程的示意图。

Notably, owing to the complex driving scenarios with uneven ground and rapid pose changes especially when turning directions to avoid obstacles during data collection, pose drifts are observed in the IMU data. Therefore, the temporal aggregation results of pointclouds are inferior with misaligned horizon and ego-motion blur as shown in Fig. A2. To relieve these issues, ICP (Interative-Closed-Point) [30] is conducted additionally for static scene points registeration before multi-frame aggregation. Finally, densified pointclouds for a single frame can be obtained by fusing the static scenes with the dynamic objects.

值得注意的是，由于数据采集过程中复杂的驾驶场景，如地面不平和快速姿态变化，尤其是在转向避障时，IMU数据中观察到姿态漂移。因此，点云的时间聚合结果较差，表现为地平线未对齐和自运动模糊，如图A2所示。为了缓解这些问题，在进行多帧聚合之前，额外进行了ICP(迭代最近点)[30]用于静态场景点的配准。最终，通过将静态场景与动态物体融合，可以获得单帧的密集点云。

Given dense points of a specific scene, we label all vox-els within a fixed range by a resolution of , based on the height of majority points inside each voxel. If the height is larger than a threshold , the voxel state is set to "occupied", otherwise "free". Moreover, considering the occlusion and truncation situations, some occupied voxels are not scanned by LiDAR beams and camera views actually. Hence we set part of voxels to "unknown" state which are invisible from both the LiDAR and camera views through tracing the casting ray.

给定特定场景的密集点云，我们根据每个体素内多数点的高度，以 的分辨率标记固定范围内的所有体素。如果高度大于阈值 ，则体素状态设置为“占用”，否则为“空闲”。此外，考虑到遮挡和截断情况，一些占用的体素实际上并未被LiDAR光束和相机视角扫描到。因此，我们通过追踪投射光线，将部分从LiDAR和相机视角都不可见的体素设置为“未知”状态。



Figure A3. Distribution of data collection scenarios in RoboSense dataset in Google Map.

图A3. RoboSense数据集中数据采集场景在Google地图中的分布。

# B.4. Metric Comparison

# B.4. 指标比较

In addition to the evaluation of detection results with the proposed matching criteria (Center-Point distance and Closest Collision-Point distance), we also provide the corresponding evaluation results using the traditional 3D IOU (Intersection-Over-Union) matching criteria for comparison, as shown in Tab. 9. It is obvious that without distance differentiation, the evaluation results of for both LiDAR-based and Camera-based methods are all in a low level, which can not reflect the objective performance and fail to satisfy the practical application requirements of the detection model. However, the proposed matching criterion is designed to measure the locating capability of closest collision points of nearby obstacles, which is more challenging and essential for low-speed driving scenarios.

除了使用提出的匹配标准(中心点距离和最近碰撞点距离)评估 检测结果外，我们还提供了使用传统3D IOU(交并比)匹配标准的相应评估结果进行比较，如表9所示。显然，在没有距离区分的情况下，基于LiDAR和基于相机的方法的 评估结果都处于较低水平，无法反映检测模型的客观性能，也无法满足实际应用需求。然而，提出的匹配标准旨在测量附近障碍物最近碰撞点的定位能力，这对于低速驾驶场景更具挑战性和必要性。

# B.5. Scene Distribution

# B.5. 场景分布

Our RoboSense dataset contains sequences, covering 6 main categories (including 22 different locations) of outdoor or semi-closed scenarios (i.e., S1-parks, S2-scenic spots, S3-squares, S4-campuses and S5-sidewalks or S6- streets). Fig. A3 illustrates the scene distributions of our collected data constructed for RoboSense dataset, which are surrounding Dishui Lake in Shanghai, China, with several markers drew in Google Map indicating the main locations performed data collection. Besides, the illustrations for each representative scenario among the collected locations are shown in Fig. A4-A9 respectively.

我们的RoboSense数据集包含 个序列，涵盖6个主要类别(包括22个不同地点)的户外或半封闭场景(即S1-公园、S2-景点、S3-广场、S4-校园和S5-人行道或S6-街道)。图A3展示了为RoboSense数据集构建的采集数据的场景分布，这些数据位于中国上海滴水湖周围，Google地图中绘制了几个标记，指示了主要的数据采集地点。此外，图A4-A9分别展示了采集地点中每个代表性场景的示意图。

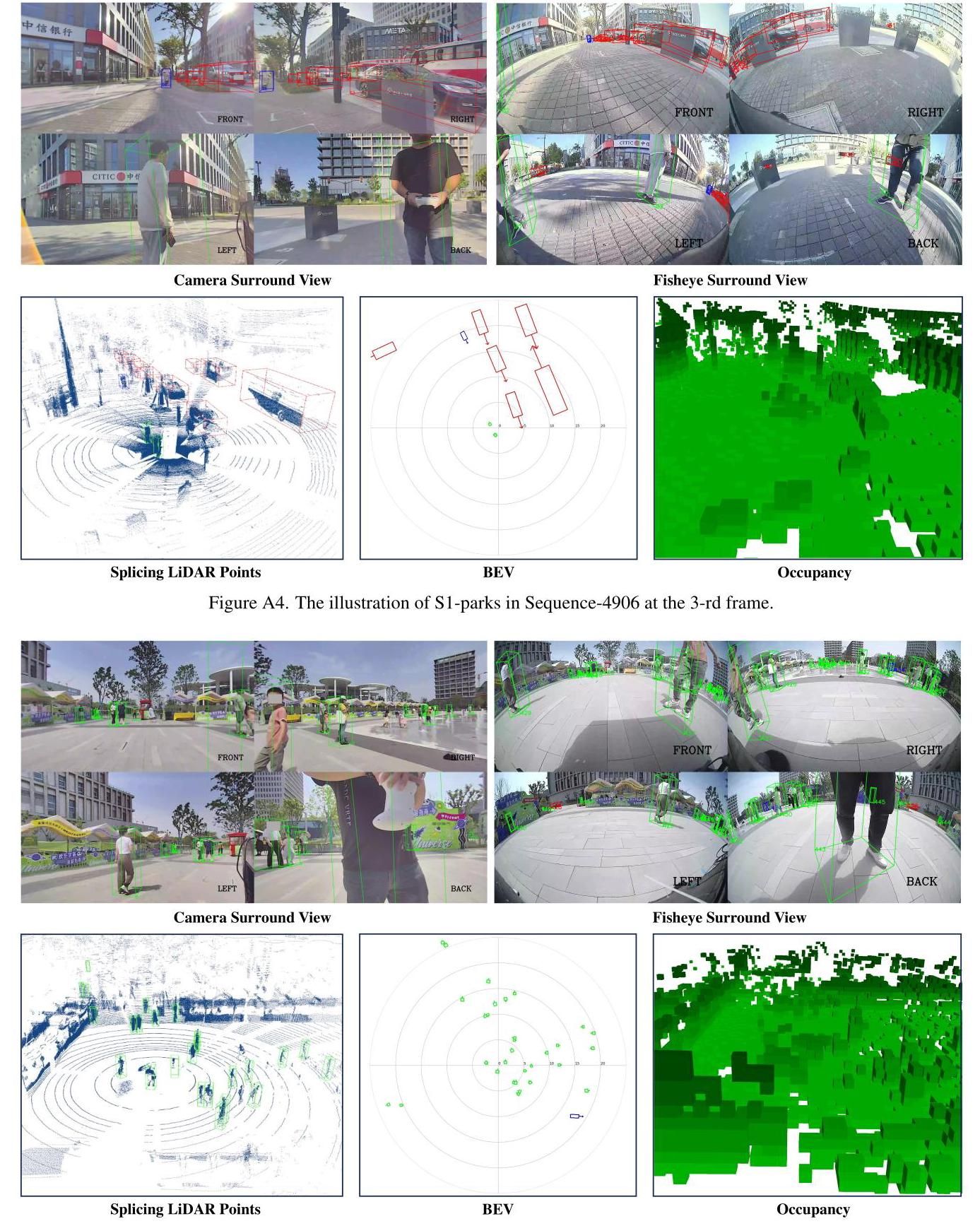


Figure A5. The illustration of S2-scenic spots in Sequence-1491 at the 13-th frame.

图A5. 序列1491中第13帧的S2-景点示意图。

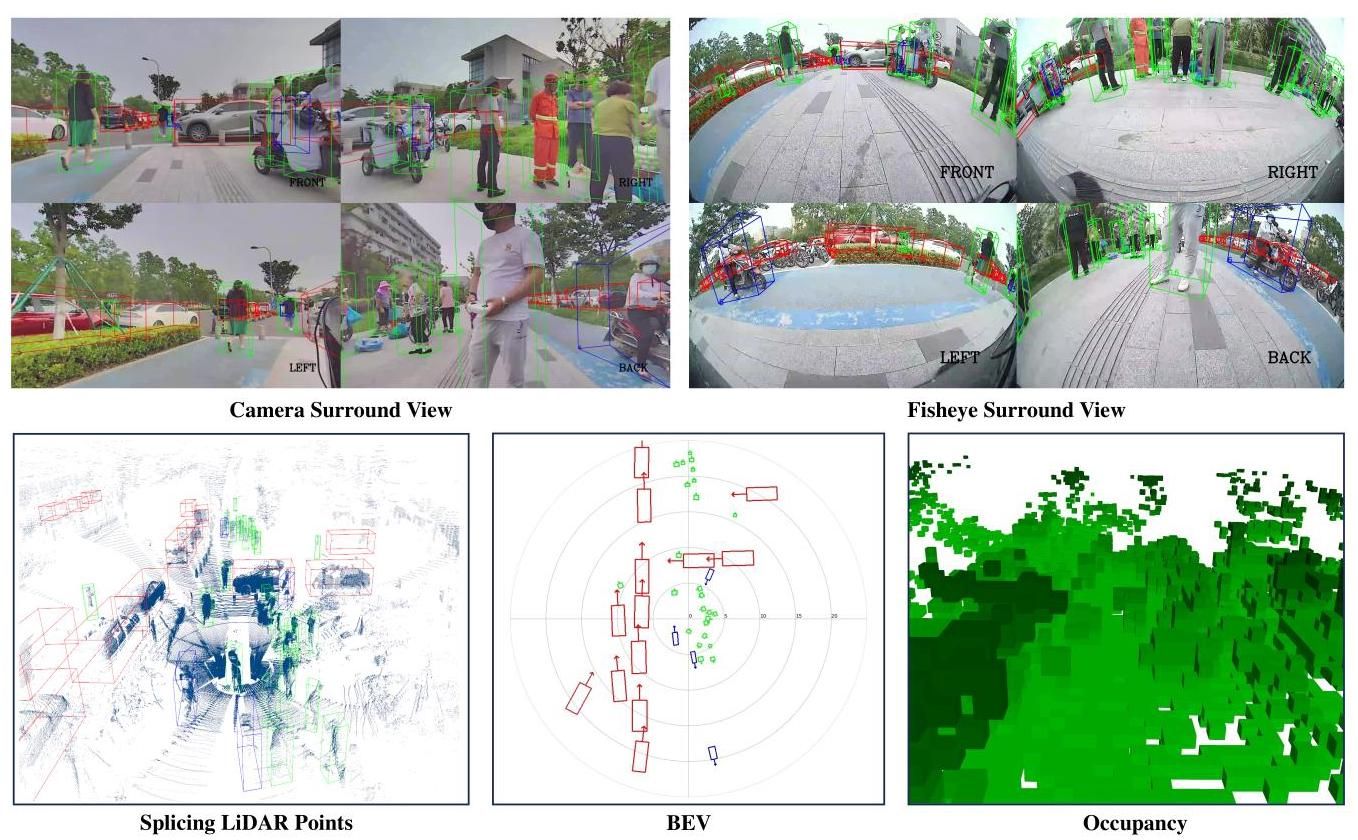


Figure A6. The illustration of S3-squares in Sequence-396 at the 2-nd frame.

图A6. 序列396中第2帧的S3-广场示意图。

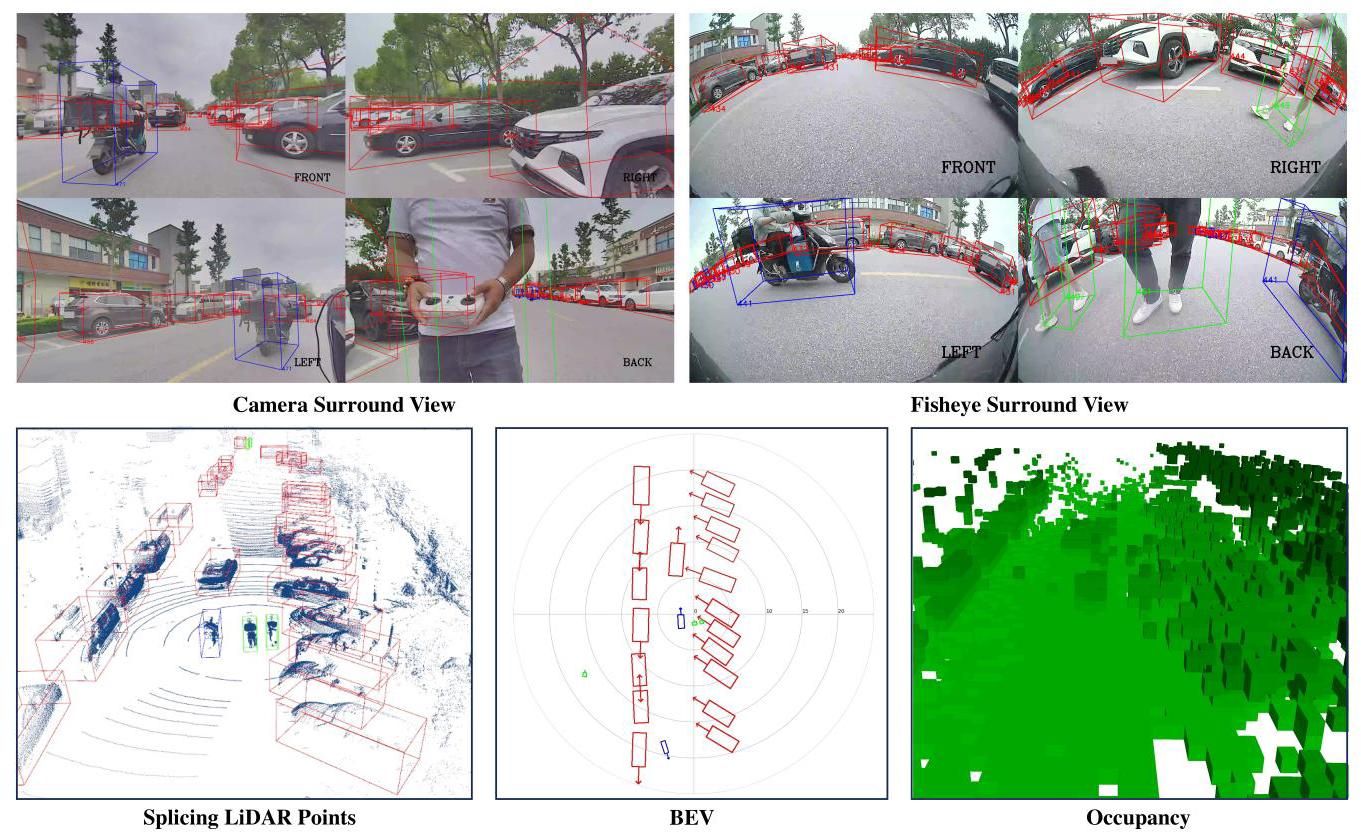


Figure A7. The illustration of S4-campuses in Sequence-2257 at the 16-th frame.

图A7. 序列2257中第16帧的S4-校园示意图。

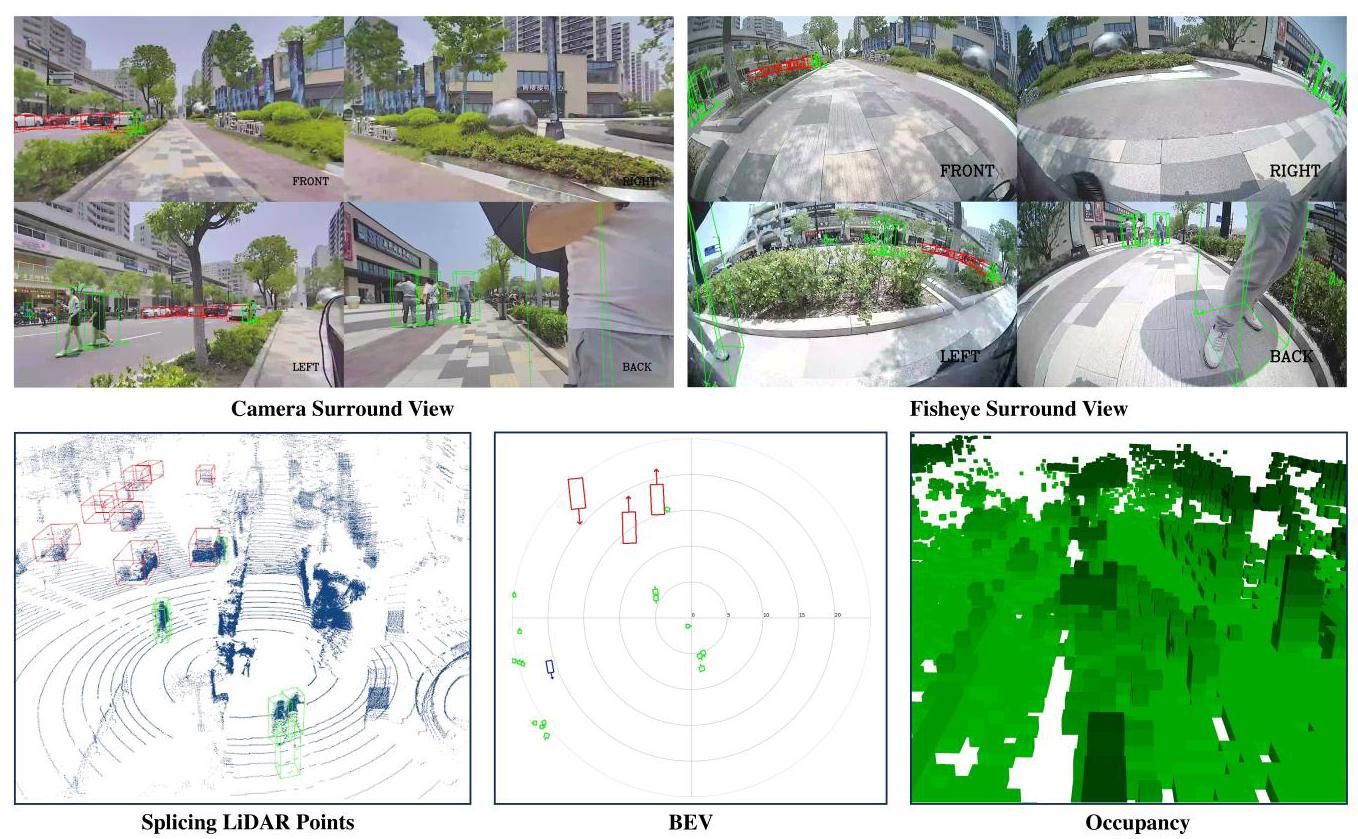


Figure A8. The illustration of S5-sidewalks in Sequence-2990 at the 10-th frame.

图A8. 序列2990中第10帧的S5-人行道示意图。

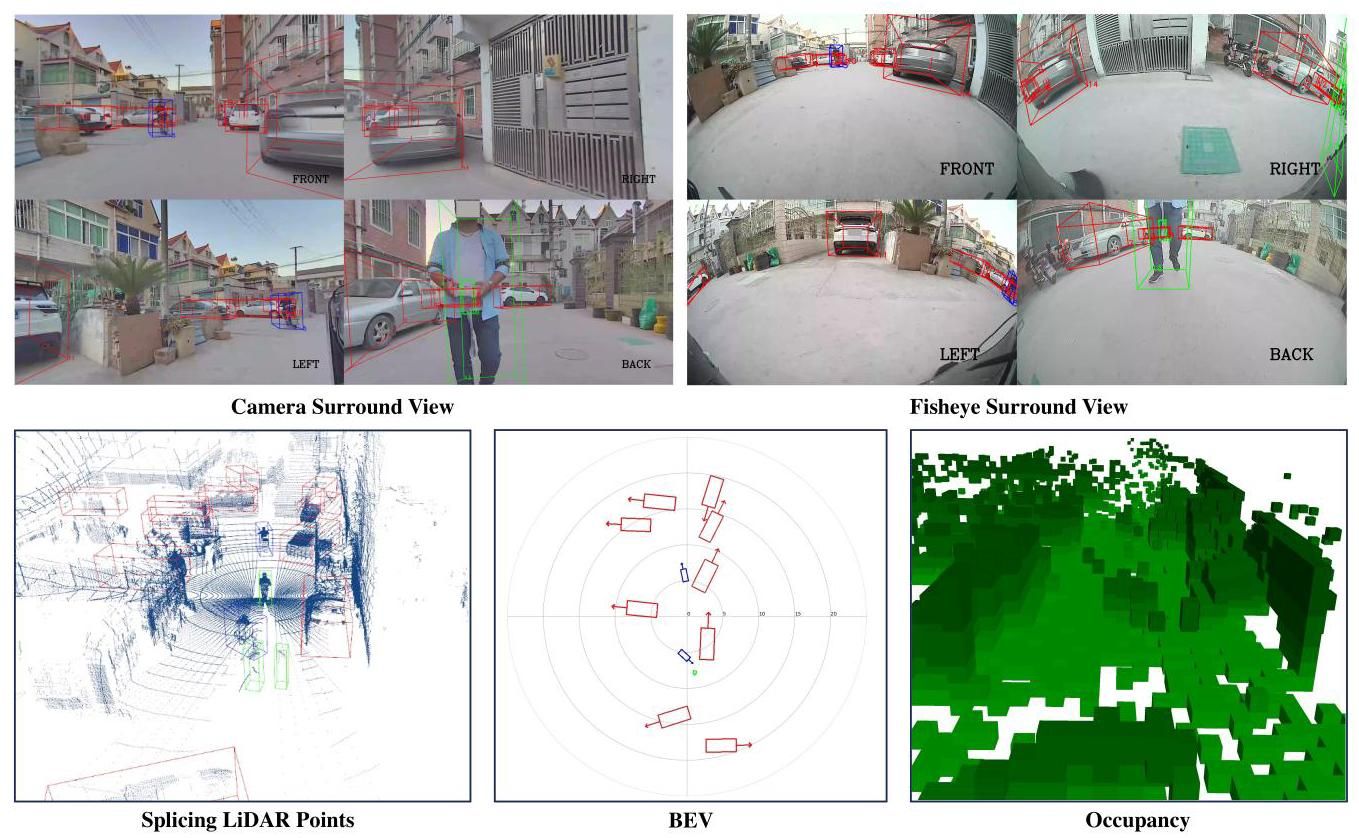


Figure A9. The illustration of S6-streets in Sequence-7018 at the 2-nd frame.

图A9. 序列7018中第2帧的S6-街道示意图。