Collaborative Perception Datasets in Autonomous Driving: A Survey

Melih Yazgan[†‡*], Mythra Varun Akkanapragada[‡*], and J. Marius Zöllner[†‡]
[†]FZI Research Center for Information Technology, Germany
[‡]Karlsruhe Institute of Technology, Germany

***Abstract*—This survey offers a comprehensive examination of collaborative perception datasets in the context of Vehicle-to-Infrastructure (V2I), Vehicle-to-Vehicle (V2V), and Vehicle-to-Everything (V2X). It highlights the latest developments in large-scale benchmarks that accelerate advancements in perception tasks for autonomous vehicles. The paper systematically analyzes a variety of datasets, comparing them based on aspects such as diversity, sensor setup, quality, public availability, and their applicability to downstream tasks. It also highlights the key challenges such as domain shift, sensor setup limitations, and gaps in dataset diversity and availability. The importance of addressing privacy and security concerns in the development of datasets is emphasized, regarding data sharing and dataset creation. The conclusion underscores the necessity for comprehensive, globally accessible datasets and collaborative efforts from both technological and research communities to overcome these challenges and fully harness the potential of autonomous driving.**

***Index Terms*—Autonomous driving, collaborative perception, dataset, V2X communication**

I. Introduction

In the evolving landscape of Intelligent Transportation Systems (ITS), there is a significant shift toward collaborative perception, which enhances the capabilities of autonomous driving and traffic management systems. Central to this shift is the implementation of V2X communications, which includes interactions such as V2V [1], V2I [2], and even Vehicle-to-Pedestrian (V2P) [3]. This advanced approach substantially improves traditional single-vehicle detection systems, offering a more comprehensive and accurate understanding of complex traffic environments. One of the key advantages of collaborative perception lies in its ability to overcome the inherent limitations of individual vehicle systems, particularly in dealing with occlusions and detecting long-range objects and sensor noise [4]. Integrating data from multiple sources increases the field of view, leading towards a holistic view of the surroundings. This multi-faceted perception enhances safety by providing a more accurate representation of the environment and contributes to more efficient traffic flow and better decision-making capabilities for autonomous vehicles. Established single-vehicle datasets such as KITTI [5], nuScenes [6], and Waymo [7] do not address the complexity of collaborative perception in addition to limitations such as sensor heterogeneity, communication protocols testing, information fusion, testing and validation of collaborative perception frameworks. Recognizing these limitations, researchers have published various datasets to test and benchmark frameworks under conditions that mimic real-world scenarios involving multiple vehicles and infrastructure components.

To the best of our knowledge, this work presents the most comprehensive collection of datasets for V2V and V2I research to date, incorporating road intersection datasets. Intersections represent some of the most complex and dynamic urban traffic environments, where various agents such as vehicles, pedestrians, and cyclists interact [8]. For autonomous vehicles, navigating through intersections poses a formidable challenge. The unpredictability and diversity of scenarios encountered at these junctions necessitate advanced perception and decision-making capabilities [9]. Furthermore, infrastructure sensors crucially enhance perception by providing vital environmental data less prone to the blind spots and occlusions typical of vehicle-mounted sensors [10]. The paper's main contributions include:

- Comparison between different datasets based on diversity, sensor setup, quality, public availability, and downstream tasks such as 3D object detection, object tracking, motion prediction, trajectory prediction, and domain adaptation.
- Comprehensive discussion on the challenges and domain gaps encountered by datasets, along with exploring the scope of future work to address these issues, is a vital aspect of this study.

Some datasets have been excluded due to their limited size [11], lack of information regarding size, annotation, and benchmark [12].

The paper is organized as follows: Section II-A systematically analyzes the road intersection datasets, presenting a comparison in Table I. This is followed by a systematic analysis of collaborative perception datasets in Section II-B with a summarized comparison in Table II. Section III discusses open challenges for future research. Section IV summarizes the key findings of this review.

II. Datasets

A. Road Intersection Datasets

Road intersection datasets are crucial, where the dataset requires diverse camera angles to capture the complex intersection traffic, and environment conditions present unique challenges [9]. These datasets are instrumental in refining 3D object detection and localization, addressing occlusions and

∗ These authors contributed equally.

Dataset	Year	Source	Sensors-(Size)	3D Labels	Classes	Tasks	Website	Public
BAAI-VANJEE [13]	2021	Real	C-(2,500), L-(5,000)	74,000	12	OD	Link	✓
IPS300+ [8]	2021	Real	C-(14,198), L-(14,198)	4,5M	7	OD	Link	✓
Rope3D [9]	2022	Real	C-(50,000)	1,5M	13	OD	Link	-
TUMTraf-I [14]	2023	Real	C-(4,800), L-(4,800)	57,406	10	OD	Link	✓
RCooper [15]	2024	Real	C-(50,000), L-(30,000)	30,000	10	OD, OT	Link	✓

[1] Sensors: Camera (C), Lidar (L)
[2] Tasks: Object Detection (OD), Object Tracking (OT)

TABLE I: Overview of Road Intersection Datasets

truncations. Table I provides the details of the datasets at a glance for further reference.

BAAI-VANJEE [13] is an infrastructure-side open-sourced dataset with highly diverse scenes. The data, which includes 2,500 frames of LiDAR data and 5,000 frames of RGB images, were recorded in sunny, cloudy, and rainy weather conditions. Its 74,000 3D and 105,000 2D object annotations distinguish the dataset. The dataset focuses on 12 classes, such as pedestrians, bicycles, motorcycles, and several types of vehicles and roadblocks. The VANJEE smart base station, approximately 4,5 meters high at a Chinese intersection, collects the data. It is equipped with a 32-channel LiDAR sensor and four cameras. Focusing mainly on tricycles and frames with densely packed object instances, this dataset features a significantly higher number of annotations per frame compared to the KITTI [5] dataset. The dataset is available under a non-commercial license.

IPS300+ [8] is multi-modal, the first large-scale open-sourced roadside perception dataset. IPS300+ covers an extensive area of 3000 square meters and extends to 300 meters. Being designed from evening rush hour scenarios in the Haidian district, Beijing, China, comprises 14,198 frames. Each frame contains an average of 319,84 labels, which is significantly higher than many existing datasets like KITTI [5]. Labeling incorporates LiDAR point clouds and images, ensuring accurate 3D bounding box annotations for categories, including pedestrians, cyclists, tricycles, cars, buses, trucks, and engineering vehicles. Each intersection in the dataset has an 80-channel LiDAR, two RGB cameras, and one GPS, providing a comprehensive view of the surrounding environment. The dataset provides a label document consistent with KITTI [5] and contains more pedestrians and vehicles per frame than KITTI [5] or nuScenes [6]. The statistics in the dataset show that the annotation size of buses and trucks is relatively smaller than that of cars, which can affect the detection accuracy of those classes. The dataset employs time synchronization and spatial calibration between different units, ensuring the consistency of labeling and accuracy of the collected multi-modal data. While the dataset's performance in 3D LiDAR detection assessed using baseline PointPillar [16], its evaluation with camera-based methods remains unexplored, emphasizing a notable gap in monocular approach assessments. The dataset is publicly available under the CC BY-NC-SA 4.0 license.

Rope3D [9] is another dataset specifically developed for monocular 3D object detection tasks. Rope3D includes a collection of 50,000 images and 1.5 million 3D annotations. Comprising 13 object classes, it provides a detailed representation of roadside elements. The primary classes include cars, oversized vehicles, pedestrians, cyclists, and extra classes such as traffic cones, triangle plates, and unknown-unmovable objects. The dataset utilizes roadside cameras mounted on poles or traffic lights and LiDAR sensors equipped on vehicles parked or moving. Rope3D comprises images captured across various lighting conditions, weather conditions, and diverse road scenes. Each has distinct camera specifications, including focal length, pitch angle, and mounted height. The annotation process focused on accurately aligning 2D and 3D annotations and paying particular attention to occlusion and truncation levels. This approach utilizes a model to bridge 3D locations and 2D projections, effectively addressing the challenge of non-parallel optical axes in cameras with varying pitch angles, thereby solving the gap as previously discussed in IPS300+ [8]. Benchmarking involves evaluating adapted monocular 3D object detection such as M3D-RPN [17]. Addressing ethical concerns, Rope3D anonymizes sensitive information such as license plates and human faces. It also restricts the use of time-discrete images to prevent potential misuse for illegal surveillance. The dataset is governed by strict usage terms outlined in a detailed confidentiality and use agreement, restricting its availability outside specified conditions and prohibiting open-source distribution.

TUMTraf-I [14] is another multi-modal large-scale open-sourced dataset. TUMTraf-I comprises 4,800 images and LiDAR point cloud frames, which include over 57,406 labeled 3D annotations. These are partitioned into ten distinct object classes of traffic participants, offering a wide range of classes in real-world scenarios. The classes include cars, trucks, trailers, vans, pedestrians, motorcycles, buses, bicycles, emergency vehicles, and others. This dataset is equipped with two cameras and two LiDARs mounted at a height of 7 meters, offering a 360° field of view. This elevated perspective is essential for observing traffic scenarios, such as turns, overtaking maneuvers, and lane merges. Accordingly, the dataset is segmented into four subsets (S1 to S4), each encapsulating different atmospheric conditions and providing a realistic spectrum of driving scenarios. Subsets S1 and S2 capture 30-second sequences during dusk, presenting continuous camera footage alongside labeled LiDAR captures. Conversely, subset S3 offers a detailed 120-second sequence shot in bright daylight, while subset S4 provides a 30-second nighttime recording amidst heavy rainfall. TUMTraf-I is used to assess

3D object detection, employing PointPillars [16] for LiDAR-based and MonoDet3D [18] for camera-based approaches. Additionally, TUMTraf-I provides a development kit. The kit supports multiple dataset formats, offering versatility and ease of integration with existing models and systems. Fig 1 is created with the development kit. The dataset is publicly available under the License CC BY-NC-ND 4.0.

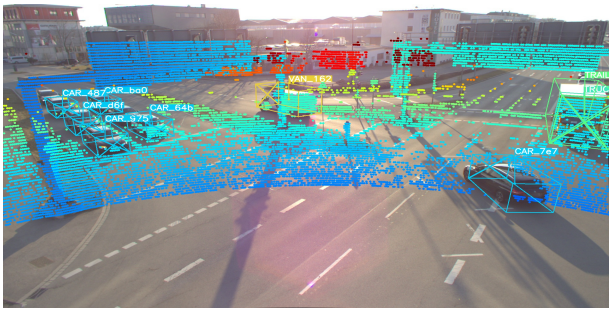

Fig. 1: 3D labels with LiDAR points on camera frame [14].

RCooper [15] is the latest real-world dataset specifically developed for roadside cooperative perception tasks. RCooper includes 50,000 images and 30,000 point clouds, covering two primary traffic scenes: intersections and corridors. This sets it apart from other datasets focused solely on intersections. RCooper provides ten object classes, which include various vehicles, cyclists, pedestrians, and construction elements. Each scene features tailored sensor setups to address specific topological challenges: intersections are equipped with a combination of MEMS LiDARs and 80-32 channel LiDARs, both operating at 10Hz, along with cameras, to capture the dynamic and congested nature of urban crossroads adeptly. Corridors are monitored with similar LiDARs and cameras, ensuring extensive coverage along extended road stretches. This varied deployment of sensors, particularly the distinction in LiDAR technologies, significantly enhances the dataset's utility for exploring challenges related to sensor heterogeneity. The benchmarking for RCooper includes evaluating cooperative perception tasks like 3D object detection and tracking using state-of-the-art methods such as [19]–[21]. The dataset is publicly available.

B. Collaborative Perception Datasets

The collaborative perception is witnessing significant advancements through the development of datasets. These datasets focus on enhancing V2V and V2X communication. By simulating complex urban environments and diverse driving scenarios, they contribute to developing algorithms for various tasks. The intricate data provided by these datasets, including detailed frame collections and comprehensive sensor setups as demonstrated in Table II, are pivotal in addressing the challenges of dynamic road conditions. For further details, reference to the official dataset pages, linked within the dataset names in Table II, is encouraged.

V2X-Sim 1.0 [22] is another open-source V2V dataset, even though the dataset's name suggests otherwise. The dataset is created by using CARLA [31] and SUMO [32]. It features a 32-channel LiDAR system with a 70-meter range, operating in a dense traffic simulation within the Town05 environment. Each scenario includes 20s traffic flow and recordings at 5Hz. Within each scene, 2-5 vehicles are randomly chosen as Connected Autonomous Vehicles (CAVs). The dataset format, derived from nuScenes [6], is extended to multi-agent scenarios, containing 10k frames in total. The dataset focuses on the 3D perception task and proposes a trainable, dynamic collaboration graph to control agent communication. Comprehensive benchmarks conducted in 3D object detection have shown that the proposed DiscoNet [22] outperforms methods such as V2VNet [1], Who2com [33], and When2com [34] in terms of the performance-bandwidth trade-off and communication latency. The dataset has been designed to be reproducible for future research.

V2X-Sim 2.0 [4] is the first open-source simulated V2X dataset, a V2I extension version of V2X-Sim 1.0 [22]. It captures traffic flow at intersections in three different CARLA towns, maintaining the same frame rate and total frame count as V2X-Sim 1.0 [22]. Each vehicle includes RGB cameras, LiDAR, GPS, and IMU, while Road Side Units (RSUs) are outfitted with RGB cameras and LiDAR as demonstrated in Fig 2. Vehicles have six RGB cameras based on the nuScenes [6] configuration, and RSUs have four cameras pointing in all directions at intersections. The dataset is benchmarked simultaneously for 3D BEV object detection, tracking, and segmentation with intermediate collaborative methods [1], [22], [33], [34]. The dataset is available under a non-commercial license. The entire dataset with V2X-Sim 1.0 [22] is open-sourced.

OPV2V [20] is another simulated open-source dataset for V2V communication, includes various roadway types and scenarios. The dataset was generated using CARLA [31] in conjunction with the OpenCDA [23] co-simulation tool, featuring multiple CAVs equipped with a comprehensive sensor setup. It comprises more than 70 scenes, 11,464 frames, and 232,913 annotated 3D vehicle bounding boxes, gathered from eight towns in CARLA and the digital town of Culver City, Los Angeles. Each frame has, on average, approximately three CAVs, with a minimum of two and a maximum of seven CAVs. Each CAV has four cameras, 64-channel LiDAR, and GPS/IMU sensors. The sensor data is streamed at 20 Hz and recorded at 10 Hz. The dataset covers frames from short scenarios in six road types: suburban midblock, urban T-intersection, urban curved road, freeway entrance ramp, urban 4-way intersection, and rural curvy road. The dataset supports collaborative 3D vehicle detection, BEV semantic segmentation, and tracking tasks only in V2V scenarios. Its benchmarking includes three fusion strategies, with the effect of CAV quantity and detection accuracy-compression trade-off. These are applied only in 3D Lidar-based object detection methods like VoxelNet [35] and PointPillar [16]. The dataset is made fully reproducible through the inclusion of driving logs. The dataset is available under a non-commercial license.

DAIR-V2X [24] is a large-scale V2I collaborative perception dataset derived from the real world. It features 71,254 LiDAR and camera frames with various vehicle types and

Dataset	Year	Source	V2X	Sensors	Size	Agents	Tasks	PA/R
V2X-Sim 1.0 [22]	2022	Sim	V2V	L	10,000	2-5	OD, OT, SS	✓/✓
V2X-Sim 2.0 [4]	2022	Sim	V2V, V2I	C, L	10,000	2-5	OD, OT, SS	✓/✓
OPV2V [23]	2022	Sim	V2V	C, L	11,464	2-7	OD, OT	✓/✓
DAIR-V2X-C [24]	2022	Real	V2I	C, L	38,845	2	OD	- / -
V2XSet [10]	2022	Sim	V2V, V2I	C, L	11,447	2-7	OD	✓/ -
DOLPHINS [25]	2023	Sim	V2V, V2I	C, L	42,736	3	OD	✓/✓
LUCOOP [26]	2023	Real	V2V	L	54,000	3	OD, OT	✓/ -
V2V4Real [27]	2023	Real	V2V	C, L	60,000	2	OD, OT, DA	✓/ -
V2X-Seq(SPD) [28]	2023	Real	V2I	C, L	15,000	2	OD, OT, TP	- / -
DeepAccident [29]	2023	Sim	V2V, V2I	C, L	57,000	1-5	OD, OT, SS, MP, DA	✓/ -
TumTraf-V2X [30]	2024	Real	V2I	C, L	7,500	2	OD, OT	✓/ -

[1] Public Availability (PA), Reusability (R)
[2] Sensors: Camera (C), Lidar (L)
[3] Tasks: Object Detection (OD), Object Tracking (OT), Semantic Segmentation (SS), Trajectory Prediction (TP), Motion Prediction (MP), Domain Adaptation (DA)

TABLE II: Overview of Collaborative Perception Datasets

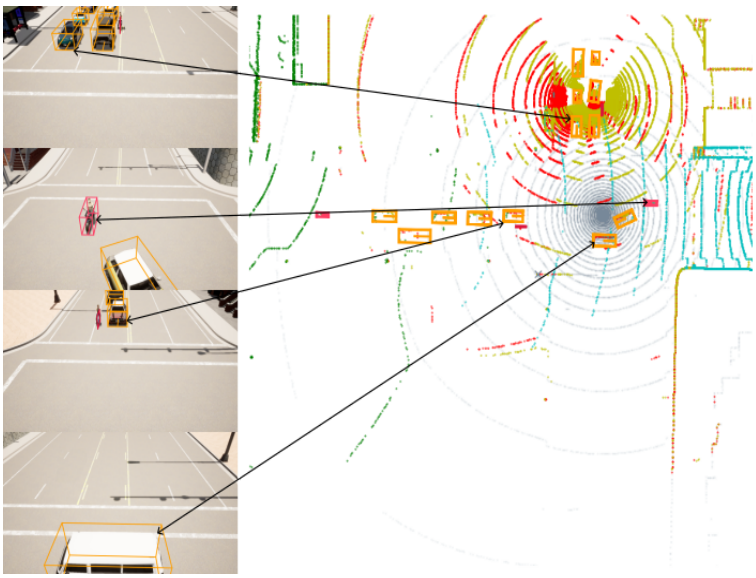

Fig. 2: The left panel shows the RSU detection frames and the right panel illustrates a LiDAR point cloud dataset, where the RSU is denoted in grey and an array of distinct colors distinguishes the various CAVs [4].

pedestrians, including cyclists and motorcyclists. The dataset encompasses various environments, including 10 kilometers of urban roadways, an equal distance on highways, and 28 distinct intersections, all captured under varying weather conditions and lighting scenarios. The dataset is divided into three main subsets, with the DAIR-V2X-C subset focusing on V2I collaboration. This subset is particularly notable for introducing the Time Compensation Late Fusion (TCLF) framework, which was developed to address the challenges of temporal asynchrony by using a specialized asynchronous subset from DAIR-V2X-C. Alongside the DAIR-V2X-C subset, the dataset also features the DAIR-V2X-V and DAIR-V2X-I subsets, focusing on vehicle and infrastructure only. Unlike others, the dataset incorporates both 3D LiDAR and image detection. For LiDAR detection, it leverages PointPillar [16] and implements both early and late fusion techniques, accommodating synchronous and asynchronous data, and includes the TCLF framework. The late fusion framework utilizes ImvoxelNet [36] as the 3D detector with synchronous data for image detection. The license conditions of this dataset mirror those of Rope3D [9], adhering to identical usage and distribution terms.

V2XSet [10] is an open-source simulation dataset that considers real-world challenges in V2X collaboration using CARLA [31] and OpenCDA [23]. It comprises 55 representative scenes covering five roadway types: straight, curvy, intersection, midblock, and entrance from eight towns. Statistical analysis shows that the dataset is biased on intersection data. Comprising 11k frames, V2XSet incorporates both V2X cooperation and realistic noise simulation, unlike DAIR-V2X [24] or OPV2V [20]. Each vehicle is equipped with 32-channel LiDAR mounted on the top and infrastructure sensors at approximately 4.5 meters, which record at 10 Hz. Each scene contains at least two and, at most, seven intelligent agents and lasts 25 seconds. The dataset is used for evaluating the effect of spatial and temporal uncertainties on 3D object detection accuracy in collaborative intermediate fusion methods such as V2VNet [1], AttFuse [20], F-Cooper

[19], and DiscoNet [22]. The dataset is released under a non-commercial license.

DOLPHINS [25] is a large-scale, open-source V2X dataset generated using CARLA [31]. It distinguishes itself from other simulation datasets by featuring dynamic weather conditions across 42,736 frames and 292,549 3D annotations compatible with the KITTI format [5]. The dataset includes at least three agents per scenario, each equipped with 64-channel LiDAR and RGB cameras, providing synchronized images and point clouds from CAVs and RSUs. DOLPHINS covers six autonomous driving scenarios: urban intersections, T-junctions, steep ramps, highways on-ramps, and uniquely mountain roads and lane merging, unlike V2X-Sim [4] datasets, which have a limited viewpoint on specific scenarios. In addition to standard labeling, the dataset is enriched with two key information types: the positions of surrounding vehicles and context-sensitive labels. These elements are vital for synchronizing perception data from various viewpoints. They cover all traffic entities within a 100-meter radius ahead and behind the ego vehicle and 40 meters on either side, providing extensive and detailed coverage. The dataset focuses on vehicle and pedestrian detection and supports 2D and 3D object detection in single-vehicle Perception. Further, it benchmarks the early fusion LiDAR 3D object detection with PointPillars [16] and MVX-Net [37]. Along with the dataset, the corresponding codes are released for flexibility and extendability of the dataset on-demand. The dataset is released under a CC BY-NC-SA 4.0 license.

LUCOOP [26] is a large scale real-world V2V dataset created by Leibniz University. It stands out from the other real-world datasets, focusing on multi-vehicle urban navigation and collaborative perception. The LUCOOP dataset encompasses over 54,000 LiDAR frames, approximately 700,000 IMU measurements, 3D map point clouds, and more than 2.5 hours of 10 Hz GNSS raw data. The dataset is gathered from three vehicles equipped with LiDAR, GNSS, IMUs, and Ultra-Wide-Band (UWB) sensors, capturing a detailed view of urban environments with narrow streets and tall buildings. Furthermore, it is enriched with a LOD2 [38] city model, enhancing its urban simulation capabilities. Integrating a stationary total station and static UWB sensors is crucial for improving the dataset's accuracy. This integration contributes over 6,000 high-precision measurements that cover more than 1 km of the vehicle's trajectory. This level of granularity and precision in ground truth verification is particularly valuable for V2V and V2X range measurements. The dataset provides further 3D bounding box annotations and precise vehicle poses but includes no benchmarking. The dataset is published with a CC BY-NC 3.0 License.

V2V4Real [27] is another large-scale, real-world, multi-model dataset for V2V perception. The dataset is collected in Columbus, Ohio, and features a diverse sensor suite on two vehicles. It has 240,000 annotated 3D bounding boxes and uniquely HDMaps across five vehicle classes captured over diverse road types, including intersections, highway ramps, and urban roads. Equipped with LiDAR, front and rear mono cameras, and GPS/IMU systems, the dataset ensures

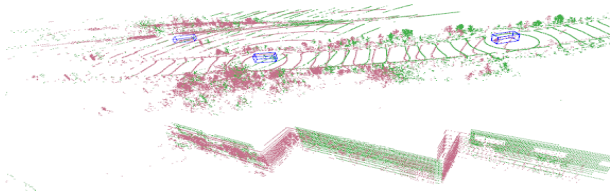

Fig. 3: Lidar point clouds, coloring relative to agents [27].

comprehensive data capture at 10Hz. The vehicles covered 410 km of road, maintaining a distance within 150 meters to guarantee overlapping sensor views as demonstrated in Fig 3. To address potential overlaps in object identification, each vehicle in the dataset is assigned a unique range of object IDs, ensuring clear differentiation. V2V4Real's benchmarking includes three fusion strategies: Late Fusion, Early Fusion, and leading intermediate methods such as AttFuse [23], F-Cooper [19], V2VNet [1], V2XVit [10], and CoBEVT [21]. These are applied over three cooperative perception tasks: 3D object detection, object tracking, and sim-to-real domain adaptation. Lidar data in OPV2V [20] and KITTI [5] format is available for download. The dataset is available under a non-commercial license.

V2X-Seq [28] is the first large-scale sequential dataset, offering data collected from real-world scenarios. Unlike DAIR-V2X [24], which focuses on 3D object detection, V2X-Seq is uniquely designed for tracking and trajectory forecasting tasks. It consists of two main parts: the Sequential Perception Dataset (SPD) and the Trajectory Forecasting Dataset (TFD). SPD, an extension of DAIR-V2X-C [24], includes over 15,000 frames from 95 scenarios, each lasting 10–20 seconds. It features vehicle and infrastructure frames sampled at 10Hz, equipped with 3D annotations for ten object classes, including unique tracking IDs for each object. TFD, on the other hand, comprises about 80,000 infrastructure-view, 50,000 vehicle-view, and 50,000 cooperative-view scenarios from 28 intersections. This subset covers 672 hours of data, providing sequences of tracked object data for 10 seconds. Additionally, the dataset includes real-time traffic light signals recorded at 10 Hz for the infrastructure portion of TFD. This data encompasses the timestamp, location, color status, shape status, and remaining time, offering significant insights into traffic participant behaviors and interactions. The V2X-seq dataset addresses challenges related to latency, and the proposed FF-Tracking method tackles the tracking task. Besides, V2X-Seq provides vector maps for intersection areas, organized similarly to Argoverse [39]. These maps contain detailed representations of lane centerlines, crosswalks, stop lines, and essential attributes like lane width and turn directions. These are crucial for building spatial context in trajectory analysis. The license conditions of this dataset mirror those of Rope3D [9], adhering to identical usage and distribution terms.

DeepAccident [29] is another large-scale open-source V2X dataset generated with CARLA [31] to represent diverse accident scenarios. It is the first simulated dataset that supports a motion prediction task. Compared to the V2X-Seq [28], it doesn't rely on precise vehicle locations, map topology, and traffic light information. The dataset features 57,000 annotated frames recorded at 10 Hz. It encompasses a variety of scenarios, including different road types, weather conditions, and times of day. The dataset's unique creation involved capturing scenes with two vehicles having overlapping planned trajectories. Additionally, two vehicles following each accident-involved vehicle and one infrastructure unit facing the intersection, summing up to five agents. Each agent has six RGB cameras and one 32-channel LiDAR. The dataset classes consist of vehicle types, including motorcycle, cyclist, and pedestrian. Specifically, DeepAccident concentrates on twelve varieties of accident scenarios at intersections, including those with and without traffic control signals. These scenarios range from running against a red light at four-way intersections to unprotected left turns and conflicting turns at three-way intersections. Besides its main focus on end-to-end motion and accident prediction, the dataset supports 3D object detection, tracking, and BEV semantic segmentation. Regarding benchmarking, the dataset's baseline model, V2XFormer, is compared against various state-of-the-art intermediate fusion modules. These include DiscoNet [22], V2X-ViT [10], and CoBEVT [21]. Finally, real-world applicability tests using the nuScenes [6] dataset reveal improved performance with models trained on both DeepAccident and nuScenes [6] data. The license conditions of the dataset are unknown, but the dataset is open-sourced.

TumTraf-V2X [30] dataset, derived from real-world data, is the latest to be released as open-source. It includes 2,000 labeled point clouds and 5,000 images, with approximately 30,000 3D bounding boxes that are enhanced with precise GPS and IMU data for accurate object location and movement tracking. Annotations conform to the ASAM OpenLABEL [40] format, and the dataset features a heterogeneous sensor setup: 32-64 channel LiDARs operating at 10 Hz and high-resolution cameras. It records a broad spectrum of traffic scenarios under various environmental conditions, including complex maneuvers such as overtaking and U-turns, and instances of traffic violations, setting it apart as a unique resource among real V2X datasets. Central to this dataset is the CoopDet3D model, a V2X cooperative perception model that utilizes vehicle and infrastructure data to improve object detection and tracking. The accompanying TUMTraf V2X development kit facilitates this data collection, providing data processing, visualization, and evaluation tools. The entire package is available under a CC BY-NC-SA 4.0 license.

III. Discussion

By comprehensively reviewing 16 collaborative perception datasets, we have identified critical areas such as domain shift, sensor setup limitations, dataset diversity, and availability. Addressing these concerns is essential for accelerating the development of autonomous driving technologies.

Domain Shift: According to Table II, it is clear that most of the datasets are created by using simulated environments, and only two of them evaluated the datasets with domain adaptation techniques. Due to inherent challenges such as labeling, privacy, and investment in gathering comprehensive real-world datasets, domain shift will likely remain an issue soon. As a result, the reliance on simulated datasets and the subsequent need for effective domain adaptation techniques are expected to be ongoing areas of focus in developing collaborative perception systems.

Sensor Setup and Limitations: As presented in Table I, the datasets are created using multiple sensor modalities. However, a critical observation from the datasets listed in Table II is the inconsistency in multi-modal approaches, especially in real-world scenarios. This indicates a significant gap in capturing the diverse and complex real-world driving conditions. Addressing limitations in vehicle and infrastructure sensors, especially under changing weather and varying light conditions, is a crucial area for further research and development. Challenges such as dealing with diverse camera angles, handling occlusions, and overcoming depth perception issues from various viewpoints are essential to address. These issues are critical for ensuring the effectiveness and reliability of data collected for both V2V and V2I applications. For V2I applications, in particular, the strategic choice of sensor heights and types necessitates further exploration. Optimizing sensor placement is key to enhancing data capture quality, which is fundamental for accurate and comprehensive environment perception. Furthermore, the future of autonomous driving, where all cars are interconnected, introduces a new layer of complexity due to the heterogeneity of sensor modalities. Car manufacturers may employ varied sensor setups, leading to diverse data types and formats. This diversity necessitates the development of effective strategies for handling and interpreting these various data.

Dataset Diversity: The study conducted by Xiang et al. [41] provides a pivotal understanding of perception in challenging scenarios, which is essential for safe and robust collaborative perception in vehicle-to-everything communication systems. These scenarios extend beyond the typical occlusions, accident prediction in 12 scenarios. Trajectory prediction will play a crucial role, especially in environments where the number of CAVs exceeds the average (see Table II). Another essential improvement is integrating Vulnerable Road Users (VRUs) into V2X communications systems, particularly V2P. Additionally, advanced tasks like anomaly detection and out-of-distribution [42] analysis are integral for evaluating and responding to unforeseen and potentially hazardous events.

Dataset Availability: The practice of open-sourcing datasets is crucial for promoting transparency and collaboration within the global research community, enabling researchers worldwide to address challenges with a shared resource pool. However, as indicated in Tables I and II of the paper, there is a noticeable disparity in the availability

of simulated versus real-world datasets. While simulated datasets are generally accessible, many comprehensive real-world datasets remain restricted, particularly in certain regions. This lack of access presents a major obstacle for researchers needing real-world data to test and advance V2X applications, regardless of location. Furthermore, the reusability of datasets, especially simulated ones, is increasingly important. Creating flexible datasets for adaptation or expansion to include specific user scenarios or tasks enhances their value and longevity. Integrating them into a unified framework like OpenCOOD [20] simplifies their application in benchmarking and comparative studies.

Privacy and Security: The development of autonomous driving relies heavily on extensive data, including images and videos from onboard and exterior cameras. Including personal details like faces, dates, and locations in this data raises concerns about privacy and security, particularly when the collected data is used for tracking or monitoring individuals without their consent. This is especially important in the creation of real-world datasets. Collaboration faces significant challenges, particularly in dataset creation, due to security concerns like malicious attacks, especially regarding sharing sensor-captured data. In response to these challenges, using Federated Vehicular Transformers, proposed by Tian et al. [43], offers a promising direction.

IV. Conclusion

In conclusion, our comprehensive overview of collaborative perception datasets has highlighted key advancements along with persistent challenges that need to be addressed. Technological progress in this area is evident, but there are notable gaps in the availability of real-world V2X datasets. Addressing these gaps is crucial for the global research community to fully realize the potential of collaborative perception. Collaborative efforts between technology innovators and the research community are essential in this endeavor. The development of extensive, globally accessible datasets will play a pivotal role in overcoming these challenges and unlocking the full capabilities of autonomous vehicles. Our review not only highlights critical gaps but also outlines a pathway for future advancements. By underlining the importance of diverse, real-world datasets and improved sensor setups, our findings encourage the development of more adaptable and robust systems. We advocate for increased collaboration and innovation in dataset creation, aiming to accelerate the progress of autonomous vehicle capabilities.

V. Acknowledgment

This work is developed within the framework of the "Shuttle2X" project, funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) and the European Union, under the funding code 19S22001B. The authors are solely responsible for the content of this publication.

References

[1] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621.

[2] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 1743–1749.

[3] C. Zhang, J. Wei, S. Qu, C. Huang, J. Dai, P. Fu, Z. Wang, and X. Li, "Implementation of a V2P-Based VRU Warning System With C-V2X Technology," *IEEE Access*, vol. 11, pp. 69 903–69 915, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10175872/

[4] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 3354–3361. [Online]. Available: http://ieeexplore.ieee.org/document/6248074/

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11 618–11 628. [Online]. Available: https://ieeexplore.ieee.org/document/9156412/

[7] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[8] H. Wang, X. Zhang, Z. Li, J. Li, K. Wang, Z. Lei, and R. Haibing, "Ips300+: a challenging multi-modal data sets for intersection perception system," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2539–2545.

[9] X. Ye, M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, and E. Ding, "Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 21 309–21 318. [Online]. Available: https://ieeexplore.ieee.org/document/9879696/

[10] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 107–124.

[11] Y. Yuan and M. Sester, "Comap: A synthetic dataset for collective multi-agent perception of autonomous driving," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 255–263, 2021.

[12] S. Busch, C. Koetsier, J. Axmann, and C. Brenner, "LUMPI: The leibniz university multi-perspective intersection dataset," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1127–1134. [Online]. Available: https://ieeexplore.ieee.org/document/9827157/

[13] D. Yongqiang, W. Dengjiang, C. Gang, M. Bing, G. Xijia, W. Yajun, L. Jianchao, F. Yanming, and L. Juanjuan, "Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china," *arXiv preprint arXiv:2105.14370*, 2021.

[14] W. Zimmer, C. Creß, H. T. Nguyen, and A. C. Knoll, "Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception," in *2023 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2023.

[15] R. Hao, S. Fan, Y. Dai, Z. Zhang, C. Li, Y. Wang, H. Yu, W. Yang, Y. Jirui, and Z. Nie, "Rcooper: A real-world large-scale dataset for roadside cooperative perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 12 689–12 697. [Online]. Available: https://ieeexplore.ieee.org/document/8954311/

[17] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.

[18] W. Zimmer, J. Birkner, M. Brucker, H. T. Nguyen, S. Petrovski, B. Wang, and A. C. Knoll, "Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023.

[19] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.

[20] J. L. J. M. Runsheng Xu Hao Xiang, Xin Xia Xu Han, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[21] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers." [Online]. Available: http://arxiv.org/abs/2207.02202

[22] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[23] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "Opencda: an open cooperative driving automation framework integrated with co-simulation," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 1155–1162.

[24] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 21 329–21 338. [Online]. Available: https://ieeexplore.ieee.org/document/9879243/

[25] R. Mao, J. Guo, Y. Jia, Y. Sun, S. Zhou, and Z. Niu, "Dolphins: Dataset for collaborative perception enabled harmonious and interconnected self-driving," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 4361–4377.

[26] J. Axmann, R. Moftizadeh, J. Su, B. Tennstedt, Q. Zou, Y. Yuan, D. Ernst, H. Alkhatib, C. Brenner, and S. Schön, "LUCOOP: Leibniz university cooperative perception and urban navigation dataset," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/10186693/

[27] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 712–13 722.

[28] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun *et al.*, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5486–5495.

[29] T. Wang, S. Kim, W. Ji, E. Xie, C. Ge, J. Chen, Z. Li, and P. Luo, "DeepAccident: A motion and accident prediction benchmark for v2x autonomous driving." [Online]. Available: http://arxiv.org/abs/2304.01168

[30] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. Knoll, "Tumtraf v2x cooperative perception dataset," *arXiv preprint arXiv:2403.01316*, 2024.

[31] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[32] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," *International journal on advances in systems and measurements*, vol. 5, no. 3&4, 2012.

[33] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.

[34] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4105–4114. [Online]. Available: https://ieeexplore.ieee.org/document/9156848/

[35] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3d object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 4490–4499. [Online]. Available: https://ieeexplore.ieee.org/document/8578570/

[36] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2397–2406.

[37] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.

[38] Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGLN), Landesvermessung und Geobasisinformation, "3D-Gebäudemodell (LoD2)," https://opengeodata.lgln.niedersachsen.de/#lod2, 2022, [Online; accessed 1-February-2024].

[39] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps. in 2019 ieee," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8740–8749.

[40] "ASAM OpenLABEL," https://www.asam.net/standards/detail/openlabel/, last accessed on 8th April 2024.

[41] H. Xiang, R. Xu, X. Xia, Z. Zheng, B. Zhou, and J. Ma, "V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3584–3591.

[42] D. Bogdoll, S. Uhlemeyer, K. Kowol, and J. M. Zöllner, "Perception Datasets for Anomaly Detection in Autonomous Driving: A Survey," in *Intelligent Vehicles Symposium (IV)*, 2023.

[43] Y. Tian, J. Wang, Y. Wang, C. Zhao, F. Yao, and X. Wang, "Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving," vol. 7, no. 3, pp. 456–465. [Online]. Available: https://ieeexplore.ieee.org/document/9857660/