A Systematic Literature Review on Vehicular Collaborative Perception—A Computer Vision Perspective

Lei Wan, Jianxin Zhao, Andreas Wiedholz, Manuel Bied, Mateus Martinez de Lucena, Abhishek Dinkar Jagtap, Andreas Festag, *Senior Member, IEEE*, Antônio Augusto Fröhlich, *Senior Member, IEEE*, Hannan Ejaz Keen, and Alexey Vinel, *Senior Member, IEEE*

***Abstract*—The effectiveness of autonomous vehicles relies on reliable perception capabilities. Despite significant advancements in artificial intelligence and sensor fusion technologies, current single-vehicle perception systems continue to encounter limitations, notably visual occlusions and limited long-range detection capabilities. Collaborative Perception (CP), enabled by Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication, has emerged as a promising solution to mitigate these issues and enhance the reliability of autonomous systems. Beyond advancements in communication, the computer vision community is increasingly focusing on improving vehicular perception through collaborative approaches. However, a systematic literature review that thoroughly examines existing work and reduces subjective bias is still lacking. Such a systematic approach helps identify research gaps, recognize common trends across studies, and inform future research directions. In response, this study follows the PRISMA 2020 guidelines and includes 106 peer-reviewed articles. These publications are analyzed based on modalities, collaboration schemes, and key perception tasks. Through a comparative analysis, this review illustrates how different methods address practical issues such as pose errors, temporal latency, communication constraints, domain shifts, heterogeneity, and adversarial attacks. Furthermore, it critically examines evaluation methodologies, highlighting a misalignment between current metrics and CP's fundamental objectives. By delving into all relevant topics in-depth, this review offers valuable insights into challenges, opportunities, and risks, serving as a reference for advancing research in vehicular collaborative perception.**

***Index Terms*—Autonomous driving, connected autonomous vehicles, cooperative-intelligent transportation systems, computer vision, collaborative perception, collective perception.**

Received 24 December 2024; revised 19 April 2025 and 15 September 2025; accepted 19 October 2025. This work was supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the program "Novel Vehicle and System Technologie" and the project "Valid Innovative Comprehensive Sensor System for Cooperative Automated Driving" (VALISENS) under Grant 19A22009E. The Associate Editor for this article was J. Fang. *(Corresponding author: Alexey Vinel.)*

Lei Wan is with XITASO GmbH, 86153 Augsburg, Germany, and also with Karlsruhe Institute of Technology (KIT), 76133 Karlsruhe, Germany (e-mail: lei.wan@partner.kit.edu).

Jianxin Zhao, Manuel Bied, and Alexey Vinel are with Karlsruhe Institute of Technology (KIT), 76133 Karlsruhe, Germany (e-mail: jianxin.zhao@kit.edu; manuel.bied@kit.edu; alexey.vinel@kit.edu).

Andreas Wiedholz and Hannan Ejaz Keen are with XITASO GmbH, 86153 Augsburg, Germany (e-mail: andreas@xitaso.com; hannan.keen@xitaso.com).

Mateus Martinez de Lucena and Antônio Augusto Fröhlich are with the Federal University of Santa Catarina (UFSC), Florianópolis 88040-900, Brazil (e-mail: lucena@lisha.ufsc.br; guto@lisha.ufsc.br).

Abhishek Dinkar Jagtap and Andreas Festag are with CARISSMA Institute of Electric, Connected and Secure Mobility (C-ECOS), Technische Hochschule Ingolstadt (THI), 85049 Ingolstadt, Germany (e-mail: abhishekdinkar.jagtap@carissma.eu; andreas.festag@carissma.eu).

Digital Object Identifier 10.1109/TITS.2025.3631141

I. Introduction

AUTONOMOUS Vehicles (AVs) are a crucial technology for intelligent transportation systems, offering the potential to significantly enhance road safety and transportation efficiency. With the emergence of Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication, Connected Autonomous Vehicles (CAVs) advance this potential by enabling data sharing not only among vehicles but also with traffic management systems, thereby adding new value to Cooperative-Intelligent Transportation Systems (C-ITS). A critical component of both AVs and CAVs is their perception capability, which involves using multiple sensors to recognize and interpret the driving environment, forming the foundation for subsequent planning and control operations. Perception tasks include 2D/3D object detection, semantic segmentation, object tracking, and motion prediction, among others. Driven by advances in artificial intelligence and multi-sensor fusion, the perception capabilities of individual vehicles have significantly improved. However, these capabilities are still limited by challenges such as visual occlusion and long-range detection, which are difficult to overcome with onboard sensors alone. These limitations can lead to reduced situational awareness, increase the risk of traffic accidents, and reduce the driving efficiency.

To address the limitations of individual vehicle perception, Collaborative Perception (CP)[1] supported by V2V and V2I communication has gained significant attention [1]. In the context of CP, where only vehicles and infrastructure are equipped with sensors, CP utilizing both V2V and V2I is

[1]In the context of collaborative perception, the terms cooperative and collective perception are frequently used. However, in this paper, we specifically use the term Collaborative Perception to emphasize the dual aspects of information sharing and coordinated action among agents. In contrast, cooperative perception focuses on information sharing, while collective perception emphasizes the distributed nature of shared perception.

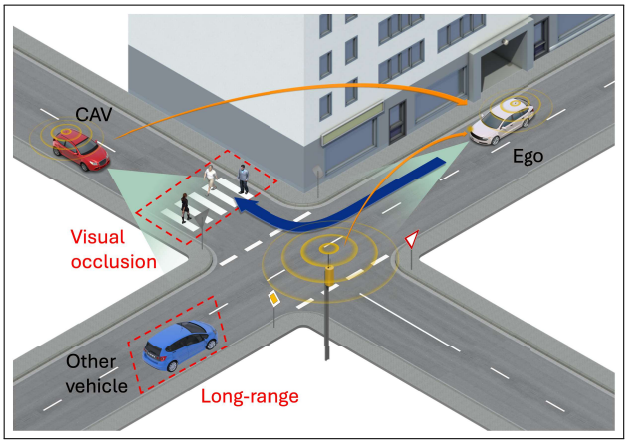

Fig. 1. Illustration of a road traffic scenario for Collaborative Perception (CP): The green shaded areas represent the ego-vehicle's (white) and the CAV's (red) Field of View (FOV). The ego vehicle cannot perceive the pedestrians on its right due to the visual occlusion caused by a building, blocking its line of sight. Additionally, another vehicle (blue) on the opposite side of the intersection lies outside the ego vehicle's perception range, presenting as the long-range problem. However, the CAV and infrastructure roadside unit can detect the pedestrians and the other vehicle, respectively, and share their observations with the ego vehicle, thereby enhancing its situational awareness.

widely described as using Vehicle-to-Everything (V2X)[2]. As shown in Figure 1, CP allows for the sharing of sensor data between vehicles and infrastructure, thereby significantly extending the Field of View (FOV) of individual vehicles to overcome challenges related to occlusion and long-range detection, which is critical for enhancing road safety and improving traffic efficiency across a wide range of use cases.

Initial investigations into CP concentrated on the transmission of object-level information [2] and aspects of the communication protocol design, such as message generation rules [3], redundancy mitigation [4] and data congestion awareness [5], and culminated in the publication of communication standards in the standard development organizations (SDOs) ETSI [6] and SAE [7]. As CP evolved, its scope broadened to include contributions from computer vision, with particular focus on the design of advanced perception algorithms and data fusion methods. Research has increasingly explored diverse data types for CP, ranging from raw sensor data [8], [9], [10], intermediate neural features [1], [11], [12], to processed perception results [13], [14], [15].

The different data types correspond to three principal paradigms of CP: early cooperation, intermediate cooperation, and late cooperation. In early cooperation, network nodes exchange raw sensor data, which contain comprehensive environmental information but require substantial bandwidth for transmission. In contrast, late cooperation involves sharing processed perception results, which is the most bandwidth-efficient data format. However, this approach is vulnerable to errors introduced during earlier perception stages, such as sensor noise, object misclassification, and data synchronization issues, and is less resilient to pose inaccuracies [16]. Intermediate cooperation is a viable solution to balance the trade-off between network bandwidth usage and accuracy. It requires less bandwidth than data-level fusion and is expected to offer higher accuracy than result-level fusion.

Each CP method offers distinct advantages and disadvantages. Nonetheless, all types consistently outperform single-vehicle perception systems that lack collaboration. CP has the potential to enhance perception accuracy and address blind spot issues. However, its practical implementation faces several significant challenges. Communication bandwidth is a significant constraint, restricting the amount of data that can be shared effectively [11], [17]. Localization errors further challenge data fusion by causing spatial misalignments [18], while time latency introduces temporal misalignments, undermining fusion accuracy [19]. Additionally, CP faces other critical challenges, including communication disruptions [20], domain shifts [21], modality heterogeneity [22], and susceptibility to adversarial attacks [23]. Overcoming these barriers is crucial for scaling CP solutions and unlocking their full potential in advancing vehicular perception systems.

A. Related Work

Several narrative reviews on CP have been published, each offering distinct perspectives on the field. For instance, Bai et al. [24] offer a high-level overview of the architecture and node structure of CP systems, while Caillot et al. [25] reviews CP, with a focus on localization, object detection and tracking. In 2023, Han et al. [26] explore CP methods for both ideal scenarios and real-world applications, highlighting the gaps between current research and practical implementation. Liu et al. [27] introduce issues of CP while Huang et al. [28] propose a generic framework of CP.

As summarized in Table I, all of these studies are narrative reviews and touch upon several aspects of CP but lack a transparent, comprehensive, and structured analysis of CP, particularly from a computer vision perspective. They do not offer a detailed taxonomy of CP technologies or fully address the range of perception tasks that benefit from collaborative approaches. For instance, key tasks such as semantic segmentation, motion prediction, and lane detection remain unexamined in prior surveys. Additionally, the role of different sensing modalities in CP has not been systematically analyzed, leaving a critical gap in understanding camera-based CP or fusion-based CP. Moreover, evaluation methodologies, which are essential for guiding the future development of CP technologies, are either absent or insufficiently discussed in previous reviews. This gap makes it difficult for readers to fully understand the range of CP tasks and to quickly identify the specific focus of their own research within the field.

To address these shortcomings, this Systematic Literature Review (SLR) follow the the PRISMA 2020 guidelines and define five research questions as below:

- **RQ1**: How can collaborative perception be classified within a structured taxonomy?
- **RQ2**: Which methodological approaches are being used for evaluating collaborative perception?
- **RQ3**: Which scenarios are covered by evaluation approaches for collaborative perception?

[2]We note that in communication technology, V2X encompasses a broader scope, covering V2V, V2I, Vehicle-to-Pedestrian (V2P) and Vehicle-to-Network (V2N).

TABLE I

SUMMARY OF SURVEYS IN VEHICULAR COLLABORATIVE PERCEPTION. MOD.: MODALITY, CO.: COLLABORATIVE TYPE, OD: OBJECT DETECTION, OT: OBJECT TRACKING, MP: MOTION PREDICTION, SS: SEMANTIC SEGMENTATION, LD: LANE DETECTION, MT/TA: MULTI-TASK.TASK AGNOSTIC, LE: LOCALIZATION ERROR, TL: TIME LATENCY, CB: COMMUNICATION BANDWIDTH CONSTRAINT, CI: COMMUNICATION INTERRUPTION, DS: DOMAIN SHIFT, HETERO.: HETEROGENEOUS SYSTEM, ADV.: ADVERSARIAL ATTACK, DA: DATASET, ES: EVALUATION SCENARIOS, EM: EVALUATION METRICS, AS: ABLATION STUDY

Paper	Year	Publication	SLR	The Taxonomy of Vehicular Collaborative Perception								Issues of Vehicular Collaborative Perception							Evaluation Method			
				Mod.	Co.	OD	OT	MP	SS	LD	MT/TA	LE	TL	CB	CI	DS	Hetero.	Adv.	DA	ES	EM	AS
[24]	2022	IEEE T-ITS			✓														✓			
[25]	2022	IEEE T-ITS				✓	✓															
[26]	2023	IEEE ITS			✓							✓	✓		✓		✓	✓	✓		✓	
[27]	2023	arXiv			✓							✓	✓	✓				✓	✓			
[28]	2024	arXiv			✓						✓	✓	✓	✓	✓	✓	✓	✓	✓			
Ours	2024	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

- **RQ4**: Which metrics are used to measure the effectiveness of collaborative perception?
- **RQ5**: What are the challenges, opportunities, and risks of collaborative perception research?

This SLR selects relevant works based on predefined inclusion and exclusion criteria and extracts key data terms from the selected papers to address the research questions. Ultimately, this review evaluates the current state of CP and highlights areas requiring further research.

B. Contributions

To minimize bias, enhance transparency, and ensure comprehensive coverage, we employ the methodology of SLR and follow the PRISMA 2020 guidelines. This review examines 106 peer-reviewed papers that meet our selection criteria, offering a summary of existing research, and a comparative analysis of critical components in cooperative perception algorithms, highlighting remaining research gaps. The key contributions of this review are as follows:

- This systematic literature review distinguishes itself from existing narrative reviews by selecting relevant works in accordance with the PRISMA 2020 guidelines, ensuring transparency and reproducibility. At the conclusion of the study, five predefined research questions are addressed.
- This review proposes a structured taxonomy for Collaborative Perception technology, addressing the limitations of prior narrow classifications in existing surveys. The taxonomy categorizes solutions along modality, collaboration and task. Furthermore, approaches to address real-world challenges in CP for autonomous driving, such as localization errors, latency, communication issues, domain shifts, heterogeneous setups, and adversarial attacks, are systematically reviewed, categorized, and comparatively analyzed.
- In contrast to the limited attention to evaluation methods in existing surveys, this review systematically examines the evaluation methodologies, performance metrics, and ablation studies employed in CP. In particular, the CP datasets are categorized and analyzed, distinguishing between synthetic and real-world datasets.
- A comparative analysis is conducted to understand the advantages and disadvantages of different methods. Building upon this analysis, the study identifies future challenges, opportunities, and risks associated with CP from various perspectives, including advancements in hardware and software for CP and improvements in evaluation methods.

C. Structure of Survey

The Sections III to VII address RQ1, beginning with an overview of a structured taxonomy in Section III. Section IV and V cover modality type and collaboration type, respectively, while Section VI explores perception tasks addressed through multi-agent collaboration. Section VII discusses the issues encountered in real-world applications and the existing solutions. Sections VIII addresses RQ2 to RQ4 and focuses on the evaluation methods of CP, with particular emphasis on the available public datasets and evaluation metrics. Section IX addresses RQ5, highlighting the challenges, opportunities, and risks in CP research. Finally, Section X summarizes the findings of the review and provides conclusions. Figure 2 provides a visual overview of the review's structure.

II. RESEARCH METHODOLOGY

A Systematic Literature Review (SLR) is a structured and methodical approach to reviewing and synthesizing existing research on a specific topic or research question. Unlike traditional narrative reviews, an SLR follows a predefined protocol that includes a comprehensive search strategy, clear criterias for selecting studies, and rigorous methods for analyzing and synthesizing the findings. The aim is to minimize bias, ensure transparency, and provide a comprehensive overview of the current state of knowledge on the topic. Our research process is following the guideline of the PRISMA 2020 statement [29] and the methodology presented in Kitchenham and Brereton [30], which serves as a transparent and uniform systematic review framework. Fig. 3 illustrates the general procedure of a SLR, which consists of three phases: Planning, Conducting, and Documenting. Additionally, the primary reviewers have diverse backgrounds in AI, computer vision, robotics, human-robot interaction, and communication, ensuring a broad range of perspectives in the review process. The application of the method to CP literature will be discussed in Section II-A, while the metadata analysis will be described in Section II-B.

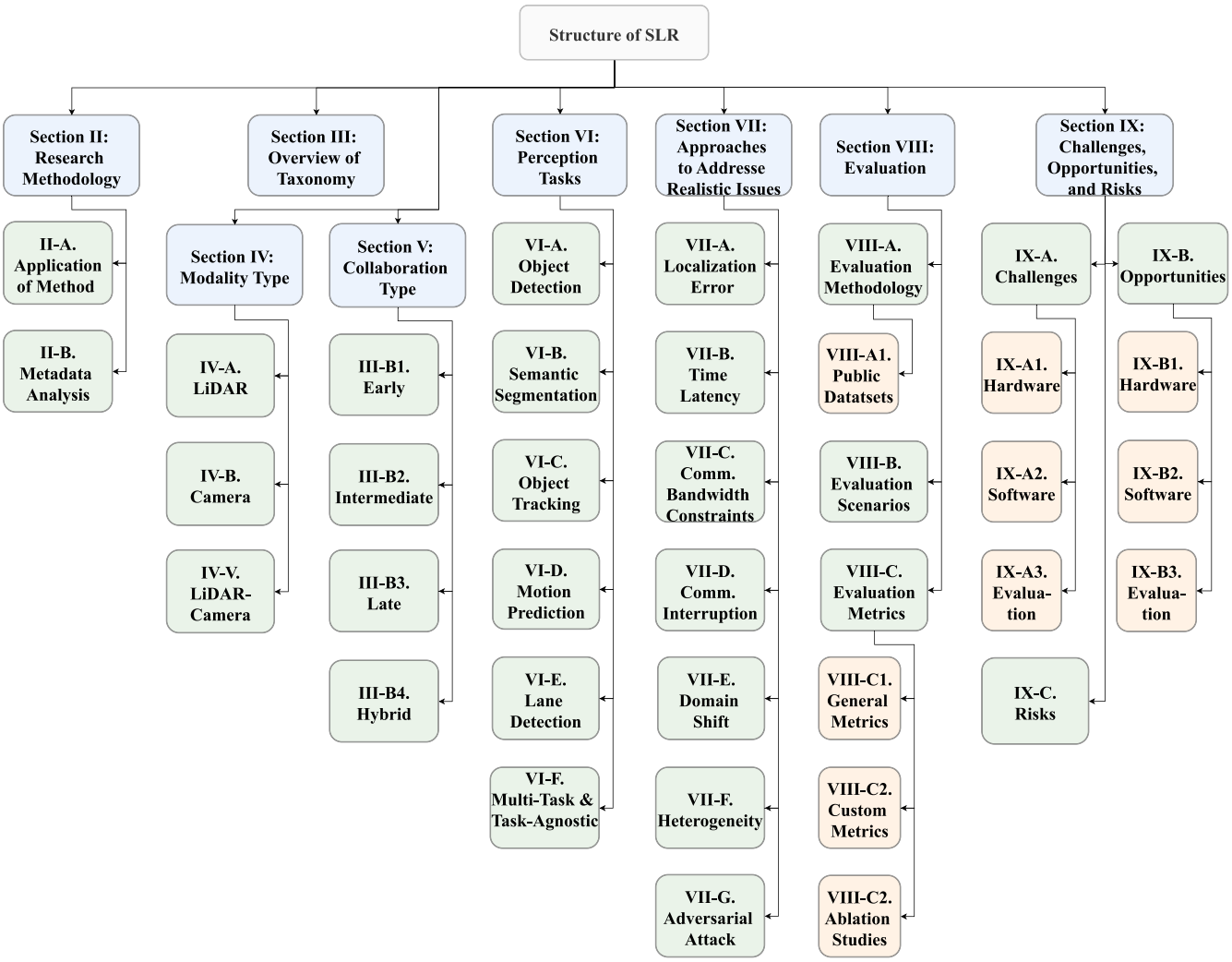

Fig. 2. Organization of this systematic literature review.

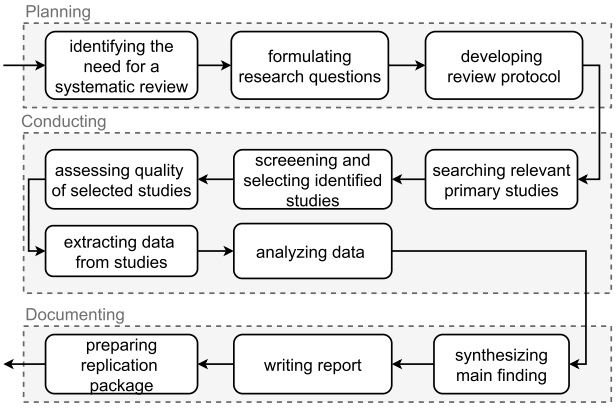

Fig. 3. The procedure of SLR in three stages: planning (review protocol development), conducting (screening and selection of articles), and documenting (synthesizing of findings).

A. Application of the SLR Method to Cooperative Perception Literature

This section will outline the practical application of the above described process. The subsequent subsections will provide a detailed explanation of each step involved in the procedure.

1) Definition of Review Protocol: The review protocol establishes the methodological framework for this study and comprises four key components: search strategy, selection criteria, data extraction strategy, and quality assurance strategy. The search strategy specifies the approach for systematically identifying relevant literature, while the selection criteria outline the criteria for including or excluding studies. The quality assurance strategy ensures the reliability of the review by evaluating the quality of the included studies.

- **Search Strategy**: The search strategy encompasses the selection of resources to be searched, the formulation of the search string, and the execution of the search procedure. In this study, several databases and one search engine were chosen as resources, as illustrated in Table II. The search string, provided in Table II, was applied across these databases and the search engine to gather relevant literature. To refine the search, appropriate filters were utilized for each resource. The main steps of the search procedure include: collecting relevant papers from

TABLE II
SEARCH RESOURCES AND SEARCH STRING

Databases and Search Engine
IEEE Xplore, ACM Library, ScienceDirect, MDPI, Scopus, Google Scholar
Search String
(collaborative OR collective OR cooperative OR multi-agent) AND perception AND (V2X OR V2V OR V2I)

TABLE III
SELECTION CRITERIA

Inclusion criteria
1) Primary studies that provide an explicit research character. 2) Studies published from 2019 to 2024 (March). 3) Studies that are classified as academic articles from conferences or journals and pre-print versions of articles that were clearly accepted by conferences or journals. 4) Studies that address (or evaluate) perception that is derived as joint effort between different entities.
Exclusion criteria
1) Studies written in any language other than the English language. 2) Grey literature that are preprints, blog posts, websites, newsletters, white papers, government documents, RSS feed, videos, podcasts and webinars, except the preprint versions of accepted papers by conferences and journals. 3) Studies that are not available, and hence not analyzable (e.g., the full text of a scientific article is not accessible). 4) Duplicates of already included studies. 5) Studies that only address the single-entity perception from vehicle perspective or infrastructure perspective. 6) Studies that do not provide details about the fusion of perception data from different entities or the evaluation of perception results in road environments. 7) Studies that only focus on the communication protocol design.

each resource up to a defined upper limit (1,000 items), removing duplicates, applying the selection criteria to the collected papers, performing forward and backward snowballing[3] on the paper set, and finally, reapplying the selection criteria.

- **Selection Criteria**: The selection criteria include both inclusion and exclusion criteria, as detailed in Table III. These criteria narrow the scope of the review to peer-reviewed academic articles published within the last five years, ensuring that the final set of papers is of high quality. Therefore, preprint papers without peer review are not included to ensure that the collected papers meet established academic standards. Specifically, exclusion criterion 6, which pertains to the level of evaluation detail, further reinforces the quality of the selected articles. The criteria are also designed to maintain a specific focus on cooperative perception techniques, explicitly excluding studies on roadside perception or ego vehicle perception.

[3]Snowballing is a technique for expanding a literature search by reviewing the references of selected papers (backward snowballing) and identifying papers that cite them (forward snowballing).

TABLE IV
DATA EXTRACTION TERM CORRESPONDING TO RESEARCH QUESTIONS RQ1

RQ1: Taxonomy of CP	
Taxonomy	Perception task, Modality/Sensor, Collaboration type, Entity type
Fusion Mechanisms	Shared information, Information fusion mechanisms, Temporal alignment mechanisms, Spatial alignment mechanisms
Repository	Repository accessibility

TABLE V
DATA EXTRACTION TERM CORRESPONDING TO RESEARCH QUESTIONS RQ2-4

RQ2-4: Evaluation of CP	
Methodology	Evaluation approach, Dataset type, Real-world experiment setup, Simulation platforms and Tools
Datasets	Supported CP tasks, V2X type, Number of CAVs, Sensor layout, Annotation, Localization of vehicle, Synchronization, Number of annotated frames, Maps, Location
Scenarios	Environment type, Road type, Traffic scenarios, Weather, Time of the day, Visual occlusion, Accident
Metrics	General metrics, Specific metrics, Ablation studies

TABLE VI
DATA EXTRACTION TERM CORRESPONDING TO RESEARCH QUESTIONS RQ5

RQ5: Challenges, opportunities, and risks of CP research	
Study	Objectives of the study, Contributions of the study, Main findings of the study, The limitation of the proposed approach, The future work of the study

An article is included in the final set only if it satisfies all the inclusion criteria and does not meet any of the exclusion criteria.

- **Data extraction strategy**: The data extraction aims to gather all relevant information necessary to address the predefined research questions. Prior to commencing the process, the specific data term to be extracted from the articles will be clearly defined and formulated. Once the final paper set is determined, the extraction strategy will be reviewed and refined to ensure both comprehensiveness and the availability of the required data. The extracted data terms are detailed in Table IV, V and VI, respectively.
- **Quality assurance**: The quality assurance process is designed to mitigate potential biases introduced by individual researchers by implementing multi-round reviews, cross-validation, and establishing consensus on key principles. The detailed quality assurance plan is outlined in Table VII.

TABLE VII
QUALITY ASSURANCE PLAN

Definition of review protocol
1) The first author defines the review protocol, including definition of research questions, search strategy, selection criteria, data extraction strategy. 2) The other authors review the review protocol 3) Disagreements will be discussed until the consensus is reached
Random assessment of included/excluded publications and extracted data
1) The first author conducts the selection/extraction process on the entire set. 2) The second author conducts the selection/extraction on a sampled subset (randomly 10%, based on the amounts of papers). 3) The outcomes of the selection/extraction on the subset are compared, and any disagreements are forwarded to the third author for discussion among the three until a consensus is achieved. 4) If the percentage of incorrectly excluded articles exceeds 10%, then it is necessary for the first author to re-examine all results considering the new consensus and return to the second step

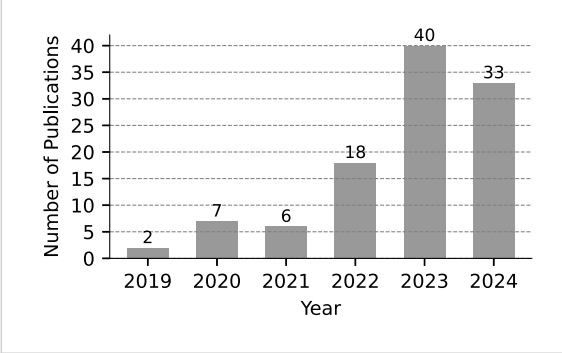

Fig. 4. Number of publications over the past five years.

2) Search and Selection: By applying the search strategy, 3,980 articles were identified after duplicate removal. The subsequent selection process involved applying the inclusion and exclusion criteria to the titles, abstracts, conclusions, and overall structure of the articles, which narrowed the paper set to 211 articles. Forward and backward snowballing techniques were then conducted on this set to ensure that no relevant articles beyond the initial search were overlooked. The selection criteria were also applied to any articles identified through snowballing. To further validate the selection, the criteria were applied to the full text of all remaining articles. This comprehensive and rigorous process ultimately resulted in a final set of 106 articles. The detailed procedure is outlined in Figure 5.

3) Data Extraction and Analysis: The data extraction strategy was initially reviewed and then systematically applied to all selected articles. The extracted data were subsequently clustered, examined, summarized, and analyzed. Both quantitative and qualitative analyses were conducted to address the research questions. These analyses enabled a clear identification of the current state of research, existing gaps, and future research trends.

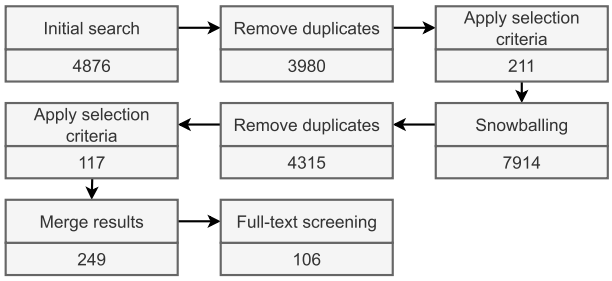

Fig. 5. Procedure of the search and selection, starting from 4876 items, reduced to 3980 after duplicates were removed, 249 after screening and snowballing, and resulting in 106 studies included in the final review.

B. Metadata Analysis

This section presents the metadata analysis conducted to identify research trends in cooperative perception. Figure 4 visualizes the number of publications from 2019 to 2024, showing a steady increase, with a pronounced surge in 2023. This trajectory reflects both the maturation of foundational research and the rapid expansion of real-world applications, supported by significant funding initiatives in intelligent transportation and autonomous driving. It also highlights that cooperative perception (CP) is transitioning from an emerging topic to a consolidated research domain. Nevertheless, due to indexing delays (e.g., IEEE Xplore updates) and the cutoff date for data collection in March 2024, the actual number of recent publications is likely higher than reported, suggesting that this upward trend is even stronger than captured here.

1) Regional Distribution: Table VIII shows that Asia (54) and North America (38) dominate the research landscape. This concentration reflects the substantial investment in V2X testbeds, 5G infrastructure, and large-scale smart mobility projects in countries such as China and the United States. Europe, while contributing fewer studies (13), plays an important role in standardization and cross-border research initiatives (e.g., ETSI standards for cooperative ITS). By contrast, contributions from other regions, including Africa (1), are minimal, underscoring the geographical imbalance of current CP research and highlighting opportunities for more globally distributed investigations.

2) Publication Venues: The majority of papers are published in premier robotics and computer vision venues, including ICRA (16), and CVPR (8) as summarized in Table VIII. This distribution indicates that CP has expanded beyond its original roots in communication and networking, and is increasingly recognized as a computer vision and AI challenge. The growing presence in CVPR, NeurIPS, and ICCV emphasizes the centrality of deep learning and visual perception methods, while robotics-oriented outlets such as ICRA and IEEE RA-L demonstrate the integration of CP into embodied autonomous systems. The shift of publication venues therefore illustrates both methodological diversification and the convergence of perception, AI, and robotics communities around CP.

3) Modalities: LiDAR-based approaches account for the majority of studies (63), with only 13 camera-only and 12 LiDAR–Camera fusion papers (Table VIII). LiDAR's

TABLE VIII
SUMMARY OF SURVEYED COLLABORATIVE PERCEPTION STUDIES ACROSS MULTIPLE DIMENSIONS

Dimension	Categories (with counts)	Key observations
Region	Asia (54), North America (38), Europe (13), Africa (1)	Research is predominantly conducted in Asia and North America.
Publication venue (top 10)	ICRA (16), CVPR (8), IEEE T-IV (8), NeurIPS (8), ICCV (5), IEEE RA-L (5), IEEE ITSC (5), IEEE T-ITS (4), IEEE IoTJ (4), IEEE IV (4)	Publications are concentrated in leading robotics and vision venues, with fewer in transportation-focused outlets.
Modality	LiDAR (63), Camera (13), LiDAR–Camera (12)	LiDAR-based approaches remain predominant, reflecting their reliability for accurate 3D detection.
Collaboration	Early (6), Intermediate (71), Late (15), Hybrid (6)	Intermediate fusion is the predominant strategy, balancing bandwidth efficiency with information richness.
Task	Object detection (78), Semantic segmentation (6), Object tracking (5), Motion prediction (3), Lane detection (2), Multi-task (3), Task-agnostic (8)	Object detection dominates the field, while other tasks remain comparatively underexplored.

predominance stems from its robustness in capturing precise 3D spatial geometry, which is critical for reliable detection in cluttered and occluded environments. Camera-based methods, by contrast, face inherent challenges in depth estimation and performance degradation under illumination changes. LiDAR–Camera fusion remains underrepresented despite its potential to combine complementary strengths (texture-rich visual data and precise depth). This limited adoption reflects the technical challenges of spatial–temporal calibration and multimodal fusion complexity. Moreover, the near absence of alternative modalities (e.g., radar, thermal, event cameras) underscores a critical research gap and signals promising directions for future multimodal CP.

4) Collaboration Strategies: Table VIII further reveals that intermediate collaboration dominates (71 studies), while early (6), late (15), and hybrid (6) approaches are far less common. The predominance of intermediate fusion reflects its ability to balance bandwidth efficiency with perception accuracy by exchanging processed features rather than raw data or final outputs. The limited adoption of other schemes underscores the persistent technical barriers: early fusion faces prohibitive communication demands, late fusion suffers from information loss, and hybrid designs increase synchronization and integration complexity. These findings suggest that while intermediate collaboration is currently the most practical solution, advancing adaptive and flexible collaboration schemes will be crucial for future large-scale deployments.

5) Tasks: Finally, research efforts are heavily skewed toward object detection (78 studies), with relatively few works on semantic segmentation (6), object tracking (5), and motion prediction (3), as shown in Table VIII. The focus on object detection underscores its central role as a prerequisite for higher-level reasoning, yet the limited attention to other tasks reveals important gaps. In particular, tracking and prediction are essential for safety-critical decision-making, and their underrepresentation highlights opportunities for future work. Similarly, multi-task and task-agnostic designs could support more integrated perception pipelines, but remain at an early stage of development.

Overall, this metadata analysis demonstrates that while cooperative perception has become a rapidly growing and increasingly multidisciplinary field, current research exhibits clear imbalances across regions, modalities, collaboration schemes, and perception tasks. These insights provide a foundation for identifying gaps and charting future research directions.

III. OVERVIEW OF STRUCTURED TAXONOMY (RQ1)

Collaborative Perception is a complex field of study with numerous subsets of sensors, collaboration methodologies and tasks. In this survey, we propose a taxonomy to classify the multitude of solutions available. We define the taxonomy based on the modality (sensor type), collaboration type, and perception task. Through the SLR, we have identified a strong focus on two types of sensor, LiDAR and camera. While the usage of LiDAR as the data source is more abundant, there is also a significant presence of cameras and the combination of LiDAR and cameras in the surveyed work. Therefore, as it relates to the modality, we classify the work into LiDAR, Camera, or a combination of LiDAR-Camera.

CP can be further classified by the collaboration type. Based on the level of the underlying data fusion algorithm, we classify work into Early, Intermediate, Late, and Hybrid Collaboration. Early, Intermediate, and Late Collaboration are self-explanatory as the data fusion inputs are shared among participants. Hybrid Collaboration refers to solutions that share data across multiple fusion levels. Furthermore, the subcategories within intermediate collaboration are outlined as follows: traditional feature fusion, attention-based feature fusion, and graph-based feature fusion.

In addition, we have identified several specific CP tasks that further classify the solutions, including object detection, object tracking, motion prediction, semantic segmentation, lane detection, multi-task approaches, and task-agnostic methods.

We further analyze the the approaches used in the surveyed studies to address realistic issues. These issues are categorized as localization errors, time latency, communication bandwidth constraints, communication interruptions, domain shifts, heterogeneity, and adversarial attacks. For each issue, we provide the corresponding categories of approaches employed to address them.

Due to the challenges associated with conducting a fair experimental comparison, such as the lack of publicly available source code for many methods, this review primarily adopts a qualitative analysis approach.

IV. Modality Type (RQ1)

In this section, we provide an in-depth examination of the different modalities in CP. Our systematic review identifies three primary modalities in the reviewed literature: LiDAR, camera, and their combination.

A. LiDAR

LiDAR is an acronym for Light Detection and Ranging. It describes a class of sensors that determine ranges by targeting an object or surface with a laser and measuring the time for the reflected light to arrive at the receiver. The sensor used in vehicular perception performs a multi-point scan across the environment at high frequencies to accurately measure the distance from the sensor to objects. The *channels* of a LiDAR sensor refer to the number of distinct laser beams emitted. It affects its resolution and field of view. For example, compared to a 16-channel LiDAR system, a 128-channel one captures more vertical slices of the surrounding environment. By varying the number of channels and their configurations, LiDAR can achieve different resolutions, ranges, and levels of detail, suitable for various applications in perception.

Just as deep neural networks can extract features from images, they can also be used to extract features from LiDAR data. One intuitive method is point-based feature extraction: process the raw data and generate a sparse representation, aggregate the features of adjacent points, and extract the feature of each point. However, this method poses stringent hardware requirements and is not seen in our surveyed work. Currently, the main feature-extraction approaches are voxel-based and pillar-based.

Voxel-based methods first convert point clouds into a structured, regular grid of 3D cells called voxels. By dividing the 3D space into voxels, the network can leverage 3D or 2D convolutional neural networks for feature extraction, making detecting objects more efficient and structured. The VoxelNet [31] is frequently used [1], [32], [33], [34], often with sparsely embedded convolutional layers applied to 3D voxel features to improve the efficiency of object detection [13], [35], [36].

The effort to improve the backbone feature extractor network is still ongoing. Besides VoxelNet, different network architectures are also proposed [37], [38]. Chen et al. [39] propose to improve the LiDAR data feature extraction backbone. They construct voxel pillars on voxel feature maps and encode them to generate Bird's Eye View (BEV) features, thereby addressing the issue of spatial feature interaction lacking in PointPillars [40] methods and enhancing the semantic information of extracted features. A maximum pooling technique reduces dimensionality and generates pseudo images, skipping complex 3D convolutional computation. In the work of Ma et al. [41], each vehicle encodes point cloud features locally using a new feature encoder network with a module called ConAda.

The pillar-based method offers advantages in real-time performance due to its efficient handling of 3D point cloud data. The pillar representation disregards partitioning along Z-axis and divides the 3D space into fixed size pillars. Intuitively the pillar is seen as an unbound voxel along the Z-axis. Pillar-based features are extracted through Deep Learning models inspired by PointNet [42]. Since pillars are not partitioned along Z-axis, a pillar-based representation of a point cloud is seen as a BEV image of multiple channels.

The pillar-based feature extractor often applied DNN on the BEV-form or raw LiDAR data. In the early phase of using this approach for LiDAR data in CP (about from 2020 to 2023), several different networks are proposed. Marvasti et al. propose such a network structure [11], [12]. Luo et al. [43] adopt the MotionNet, quantizing the 3D points into regular voxels and representing the 3D voxel lattice as a 2D pseudo-image, with the height dimension corresponding to image channels. Qiao and Zulkernine [44] use PointNet instead. The DiscoNet proposed by Li et al. [45] is later used by others [19], [46].

The most representative module for pillar-based feature extraction is PointPillars [40]. It employs a simplified version of PointNet within each pillar to extract features from the points. The point-wise features are then aggregated to create a single feature vector for each pillar. These pillar features are organized into a 2D grid, allowing leveraging 2D CNN for feature extraction.

PointPillars is widely used for LiDAR data feature extraction [9], [10], [16], [20], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64]. Some studies build on the PointPillars framework by developing structurally similar models that adapt its core principles without directly replicating it [65], [66]. Wang et al. [67] retain the PointPillars architecture but enhance it by replacing the 2D backbone with a four-layer residual network and adding a spatial pyramid pooling module. This enhancements expand the model's input area and enable it to combine information from multiple scales.

Some of the most recent research efforts try to improve the feature extraction mechanism. Instead of using the standard backbones, Bai et al. [68] propose a new adaptive feature encoder named Pillar Attention Encoder, which extracts the feature data based on the attention mechanism and adaptively reduces the data amount for sharing based on the exact communication bandwidth.

B. Camera

Cameras are among the most widely utilized modalities in perception systems, valued for their ability to capture high-resolution visual data containing dense semantic information, which is essential for tasks such as object detection, lane detection, and scene understanding. Monocular and multi-view camera setups are the two most common configurations employed in visual perception systems. Camera-only 3D perception provides an economical alternative to LiDAR-based systems. However, accurately estimating depth remains challenging due to the lack of direct 3D measurements. Similarly,

camera-only CP remains relatively under-explored, encountering challenges similar to those in single-vehicle camera-only 3D perception. Due to that there are a limited number of camera-based papers, we will introduce them separately in this part.

Hu et al. [69] introduce CoCa3D, the camera-only 3D detection improved by introducing multi-agent collaborations, while many previous work focus on network designs. The proposed CoCa3D method first enhances image-based single-agent depth estimation before the Collaborative detection feature learning module that enhances 3D detection. In the later phase, the BEV features that may contain the most informative cues are exchanged and fused to get a better BEV feature map.

Huang et al. [70] aims to achieve scalable camera-based collaborative perception with a Transformer-based architecture. The image information of the vehicles is projected into features using a BEV encoder backbone such as BEVFormer. The transformer is trained to take the BEV feature of the ego-vehicle and the poses of a collaborator and its cameras as input, and it chooses which part of the collaborator's feature map is important and should be transmitted.

Wang et al. [71] propose to address the information loss and pose errors due to time asynchrony across cameras in image-based fusion. Thus, it proposes a new fusion network architecture. It contains an attention and channel masking mechanism to enhance infrastructure and vehicle features at scale, spatial, and channel levels to correct the pose error introduced by camera asynchrony. It also uses feature compression to improve transmission efficiency. The proposed structure uses ResNet-50 as a backbone and FPN as a 2D neck to extract image features. Its evaluation is based on the DAIR-V2X dataset.

Fan et al. [72] propose the query cooperation paradigm for cooperative perception tasks, which is more interpretable than scene-level feature cooperation. They then propose the transformer-based QUEST framework utilizing VoVNetV2 [73] as the feature encoding backbone. Every query output from the decoder corresponds to a possible detected object, and the query will be shared if its confidence score meets the request agent's requirements. As the cross-agent queries arrive, they are utilized for query fusion and implementation.

C. LiDAR-Camera

Most work on CP that utilizes both LiDAR and Camera sensors follows a simple paradigm. The proposed structure can use either LiDAR or Camera data as inputs because both types of sensor data will be turned into the same type of BEV feature maps as a uniform intermediate representation for later processing. The work of Yin et al. [74] is a typical example. It proposes V2VFormer++, where individual camera-LiDAR representation is incorporated with dynamic channel fusion (DCF) at BEV space, and ego-centric BEV maps from adjacent vehicles are aggregated by a global-local transformer module. The camera images are first cropped with a resolution of 520×520 pixels, fed into the ResNet-34 encoder for multi-scale feature extraction, and then processed by a sparse cross-attention view Transformer module. PointPillars first processes the single-vehicle LiDAR data for point feature extraction, and a simple PointNet architecture is used for pillar feature extraction. Finally, a 2D CNN backbone is introduced to merge multi-resolution maps into a dense LiDAR BEV feature. Many other work follow the same pattern [22], [75], [76], [77]. These work may vary slightly in the backbone used, especially for processing camera data. For example, Zhou et al. [78] uses the Fast-SCNN network as the image feature map encoder, while some may use BEVFormer. Zhang et al. [79] provide a slightly different scenario where each agent is equipped with LiDAR and camera sensors. The work of Zhang et al. [80] fuse LiDAR and RGB data through point cloud fusion, first converting RGB images into virtual point clouds and then combining them with real point clouds.

D. Comparative Analysis

One noticeable observation is that while there is extensive research on LiDAR, camera-based CP has only recently emerged, with relatively few papers exploring the use of cameras as a data source. The reason could be the lack of depth perception of cameras, sensitivity to lighting and weather conditions, and heavy computational requirements in semantic understanding. Additionally, visual data from cameras raises privacy concerns under data protection laws, further impacting the deployment of camera-based systems. However, despite such differences, we can still observe that the problem to solve in both cases are similar, such as the limited bandwidth, lossy communication, temporal- and spatial-asynchrony, sensor and model heterogeneity, etc. They are still being actively investigated regardless of sensor types. Thus, the improvement in one area can also have an impact on the other. Besides, one trend we can observe is that research tends to use existing backbones and datasets, gradually convergent to a limited number of choices.

V. Collaboration Type (RQ1)

This section presents an in-depth review of the various collaboration types in CP: Early, Intermediate, Late, and Hybrid.

A. Early Collaboration

In collaborative perception, early collaboration refers to the approach where raw sensor data (such as camera images, or LiDAR point clouds) from multiple vehicles are shared and fused early in the processing pipeline. This is done before any significant local processing or feature extraction is applied to the data. The fused data is then processed collectively to generate a unified perception of the environment. It allows for richer information exchange, as the original details in the sensor data are preserved. On the other hand, sharing raw sensor data, such as high-resolution camera images or dense LiDAR point clouds, requires significant communication bandwidth. Some examples of this approach exist where raw LiDAR data is shared among vehicles [8], [9], [10], [81] and one where infrastructure also participates [10]. [78] differs from the others in that it also enables the sharing of raw camera data.

B. Intermediate Collaboration

In intermediate collaboration, neural network-generated features are distributed and merged to improve perception performance and conserve bandwidth. Based on their fusion mechanisms, these methods are categorized into three types: traditional feature fusion, attention-based feature fusion, and graph-based feature fusion. This section presents a comparative analysis of these intermediate fusion approaches.

1) Traditional Feature Fusion: Non-parametric operators such as summation, maximum, and average are commonly employed in neural network architectures to integrate information. These operators are particularly effective for merging features with spatial characteristics from different agents. For example, Marvasti et al. utilize non-parametric element-wise summation to fuse BEV features from multiple sources [11], ensuring comprehensive inclusion of available data. However, features with larger magnitudes may disproportionately affect the outcome, potentially overshadowing smaller yet significant inputs. Guo et al. [33] introduce a lightweight feature-based CP framework employing the maxout operator, which excels in emphasizing the most critical features or activations while being robust against variations in the number of contributing agents. Despite its effectiveness, the maximum operator risks discarding valuable contextual information by focusing solely on the highest values. Non-parametric operations are favored for their computational efficiency and simplicity of implementation. In contrast, parametric operators involve learnable parameters within the fusion module, such as convolution layers, offering a more adaptive approach to feature integration. Qiao and Zulkernine [44] propose an adaptive feature fusion model that combines spatial and channel-wise feature fusion, leveraging both max and average pooling and trainable neural layers to enhance feature extraction selectively. Another prominent method is feature concatenation followed by a trainable neural layer, as demonstrated by Bai's feature fusion backbone [37] using a dense CNN network to process concatenated features. This approach allows for the extraction of relevant information, significantly enhancing performance, though it may increase the feature dimensionality and computational demand.

To conclude, traditional feature fusion techniques utilize reduction operators and often integrate trainable neural layers to extract the most relevant features effectively. This approach strives to balance performance improvement with computational efficiency, ensuring an optimal feature integration process.

2) Attention-Based Feature Fusion: The attention mechanism [82] is effective in capturing long-range dependencies and contextual relationships, making it highly suitable for feature weighting during fusion. For example, Wang et al. propose the F-Transformer [47], a point cloud fusion transformer that employs only Transformer encoder to fuse features from different views. As illustrated in Fig. 7a, features from multiple entities are represented as tokens and forwarded to the attention module, where contextual relationships are learned to produce a fused feature representation. Unlike conventional Transformers, it omits position embeddings since the spatial arrangement of views is arbitrary, and there is no inherent ordering relationship between features from multiple perspectives. This design enhances robustness by preventing the model from making erroneous spatial assumptions and instead allowing it to focus on learning meaningful feature correlations across views. Xu et al. introduce the V2X-ViT [16], designed to fuse information across on-road agents efficiently. It enhances self-attention by incorporating an additional weight matrix tailored to the type of the source and target agents. For example, agents are categorized as either vehicles or infrastructure, and the weight matrix dynamically adjusts to optimize collaboration based on their type. Hu et al. [75] introduce a spatial confidence-aware attentive fusion, where a spatial confidence map identifies perceptual uncertainty across different areas, serving as a basis for attention learning. This method prioritizes features with higher confidence during fusion, enhancing reliability. Lu et al. [55] propose a robust multiscale attentive fusion to mitigate noise from spatial misalignment. This method leverages features at different scales: finer scales provide detailed semantic information, while coarser scales offer robustness against spatial noise, thus maintaining semantic density and enhancing overall robustness. Yang et al. [52] address temporal noise using the spatial-temporal collaboration transformer (STCFormer), which features decoupled spatial and temporal cross-attention. STCFormer follows the architecture of a vanilla transformer but incorporates three customized modules: temporal cross-attention, decoupled spatial attention, and adaptive late fusion. The temporal cross-attention captures historical context across agents to enhance the representation of the current frame, mitigating point cloud sparsity caused by fast-moving objects. The decoupled spatial attention fuses spatial features from multiple agents, while the adaptive late fusion module integrates spatial features using weight maps. With these customized modules, STCFormer achieves robust detection performance even in dynamic environments. Despite its effectiveness, the computational complexity of attention mechanisms $O(N^2)$ poses scalability challenges. To address this, Yang et al. [56] utilized a deformable cross-attention module that selectively focuses on informative locations, significantly reducing computational demands and memory usage. Unlike standard attention, which assigns weights to all elements in the feature space, deformable attention selectively attends to a sparse set of informative locations, improving computational efficiency and scalability.

LiDAR-based features inherently possess spatial characteristics suitable for per-location fusion via attention. Expanding cooperative perception to camera sensors, the BEV feature is commonly used. However, due to the inherent uncertainty in depth estimation, visual BEV features are less reliable than LiDAR features. To mitigate spatial misalignment, Huang et al. [70] propose a camera-based collaborative BEV feature fusion using selective deformable attention, which fuses features based on an interest score threshold, emphasizing relevant and significant features for ego's perception. The interest score is generated by a simple network that processes the BEV features as input, and during inference, only those features with scores above a threshold of 1 are selected.

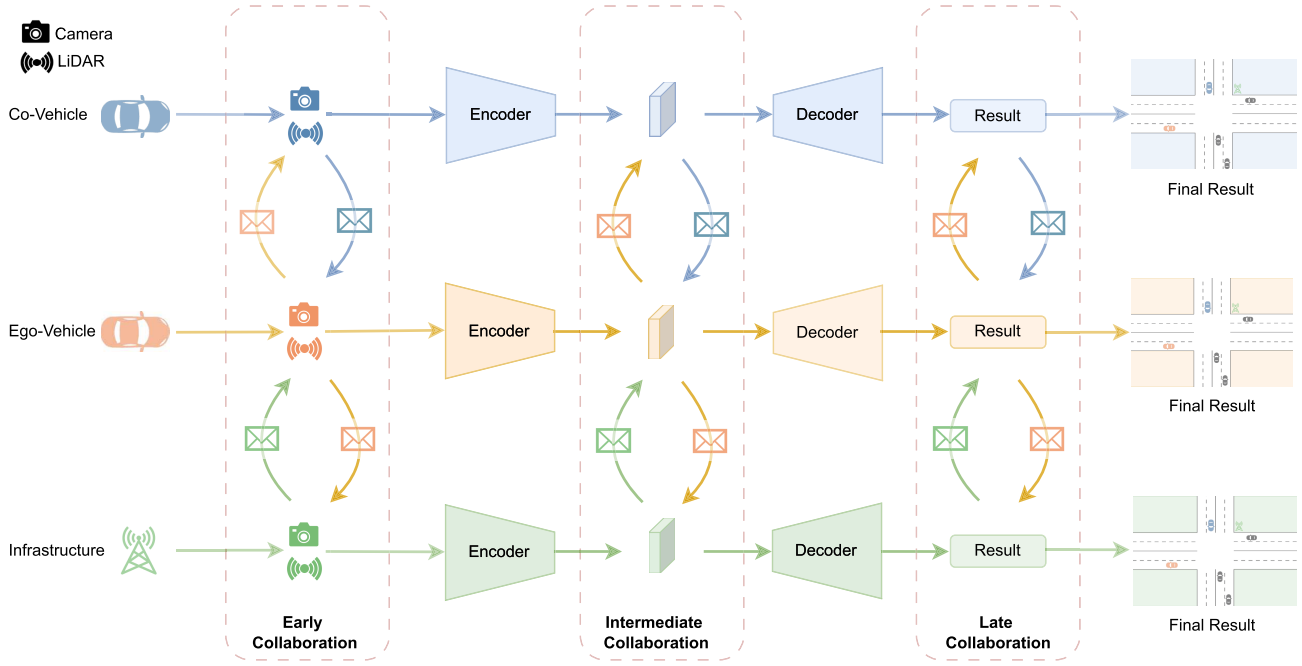

Fig. 6. Illustration of collaboration type in CP, showcasing Early Collaboration, Intermediate Collaboration, and Late Collaboration across Co-Vehicle, Ego-Vehicle, and Infrastructure, integrating Camera and LiDAR data. Early Collaboration involves sharing raw sensor data (e.g., images, LiDAR point clouds) for joint processing, while Intermediate Collaboration transmits extracted features (e.g., key points, feature maps). Late Collaboration shares final perception results (e.g., detected objects, trajectories).

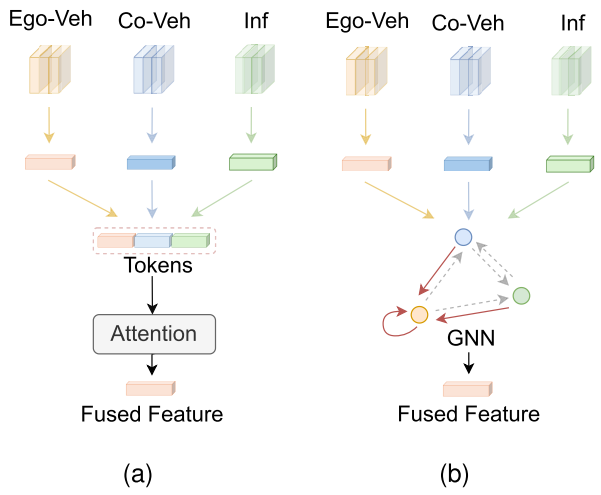

Fig. 7. Illustration of intermediate feature fusion: (a) Attention-based feature fusion, where features from ego-vehicle, cooperative vehicle, and infrastructure are transformed into tokens and aggregated using attention mechanisms; (b) Graph-based feature fusion, where features are represented as nodes and fused through message passing and node state updates in a GNN.

In conclusion, Attention mechanisms and their variants play a pivotal role in collaborative feature fusion, enhancing feature integration across channel, spatial, and temporal dimensions. Techniques like confidence mapping and deformable attention are employed to improve fusion robustness and effectiveness further.

3) Graph-Based Feature Fusion: Graph structures are practical tools to represent complex relationships among data elements, where features are modeled as nodes interconnected by edges. These edges depict interactions between features. As shown in Fig. 7b, multi-agent collaboration can be conceptualized as a graph where nodes represent individual agents and edges represent inter-agent collaborations. Graph Neural Network (GNN) [83] are well suited for processing such graph-structured data, enabling effective message passing and node state updates that facilitate information aggregation and propagation across the network. For instance, Wang et al. [84] introduce a spatially-aware GNN where each agent maintains a local graph with nodes holding state representations. These states are updated via a trainable neural network such as ConvGRU, which processes edge-weighted feature maps from all nodes to output updated node representations. This method incorporates historical context, enhancing temporal alignment and enabling joint object detection and motion prediction. Li et al. [45] propose a collaboration graph with trainable edge weights reflecting the collaboration strength between agents. These spatially and temporally aware weights allow agents to identify regions requiring collaboration dynamically. Xiang et al. [22] further evolve this concept by introducing the H3GAT, a heterogeneous 3D graph attention model that integrates attention mechanisms with GNNs. This model captures local and global interactions, preserving detail and providing a comprehensive context. Liu et al. [58] employ a multi-scale graph-attention technique to extract more comprehensive semantic information across different levels of granularity, enhancing feature integration.

In conclusion, GNNs represent a sophisticated approach to modeling multi-agent collaboration. GNN fusion with attention mechanisms enables a nuanced capture of local and global contexts, facilitating a more detailed and integrated feature analysis.

C. Late Collaboration

Unlike the previous two approaches, each agent processes its sensor data independently in late collaboration and extracts relevant results or information. The processed information (often in the form of high-level results such as object-level data) is then shared with other agents or a central system, and the fusion happens at a higher, more abstract level. Since only high-level, compact information is shared, late collaboration requires much less communication bandwidth compared to the other two approaches. Besides, it allows different agents to have varying sensor capabilities and still collaborate since only the abstract results are shared. However, some detail or precision may be lost during local processing.

The late collaboration approach is potentially modality agnostic, and in specific papers LiDAR data are often used as example [13], [36], [85]. Yu et al. [86] utilize both LiDAR and camera sensors.

A simple approach in late fusion is to use detected bounding boxes from multiple vehicles and weight them according to the detection confidence, such as Non-Maximum Suppression (NMS). Another way of late fusion is to use an adapted Kalman filter. The collectively perceived tracks are considered as measurements and integrated into the local environmental model. In general, late fusion approaches utilizes less information than the previous two approaches, and existing work's focus is mostly on improving fusion accuracy, given constraints such as heterogeneous density, low quality object proposals, overconfidence, etc.

According to the data exchanged among vehicles for cooperative object detection, the most common form is the detected object list. Zhang et al. develop a three-stage fusion scheme: partitioning local objects, generating global objects, and eliminating overlapped boxes [13]. Yu et al propose a detection boxes fusion network for the late fusion, the inputs of which are vehicle-side and road-side boxes. This network performs coordination transform, filtering, object match, and combination [86]. In [23], the authors assumes the existence of attackers and propose an approach where each vehicle samples a subset of teammates and compares the results with and without the sampled teammates. Only after a consensus is verified, indicating no attackers among the participants, that the vehicle can output the perceptual results. Sampling ensures the scalability of this solution. Teufel et al. propose to incorporate collectively detected objects to enhance the local perception capabilities [38]. Song et al. uses optimal transport theory to correct inaccurate vehicle location and heading measurements using only object-level bounding boxes [15]. Xu et al. propose mechanisms that considers confidence scores and mitigate the misalignment in box aggregation [85]. In [14], the aim is to check perceived information for its trustworthiness and validity, so other information, such as covariance information, is also exchanged.

D. Hybrid Collaboration

Yuan et al. [87] combine late and intermediate collaboration. The fusion step combines multiple types of information: object box proposals (as in late collaboration), sensor pose, selected key point coordinates, and selected features (instead of all deep features as in intermediate collaboration). The aim is to reduce the redundancy of shared deep features to decrease the communication overhead.

Wang et al. [65] employ a two-stage fusion approach. In the first stage, an edge device collects and fuses the encoded Pillar features from the LiDAR data of all cooperative vehicles to generate a list of detected objects. This object list is then transmitted to the ego vehicle, which performs a late fusion by combining it with its own object list predictions.

Dao et al. [60] propose a "late-early" collaboration framework for V2X cooperative perception. Here, objects detected by each connected agent at a past time closest to the present are broadcast. Detected objects shared between agents are propagated to the present timestamp using their velocities, computed by pooling point-wise scene flow. These propagated objects are then fused with the point cloud collected by the ego vehicle at the current time to enhance its perception. This work relaxes the assumption of inter-agent synchronization to agents sharing a shared time reference (e.g., GPS time) and acknowledges that agents produce detections at different rates. As a result, exchanged detections always have older timestamps than the timestamp of the query made by the ego vehicle, thus risking a misalignment between exchanged detections and their associated ground truth. To resolve this issue, the method simultaneously predicts both the velocities and locations of objects by pooling point-wise scene flow, effectively correcting for temporal discrepancies.

Liu et al. [64] also combine intermediate and late collaboration approaches. In the proposed fusion scheme, LiDAR data is divided into two types according to the overlapping area between the detection ranges of vehicles. For the overlapping area, intermediate collaboration is applied by sharing and fusing the features from different vehicles. For the non-overlapping area, late collaboration is conducted by generating and sharing the local detection result with an economic bandwidth.

Xie et al. [35] combine all fusion approaches. This framework enables vehicles to partition each point cloud frame into three parts: raw, feature, and object data, and exchange the data with other vehicles. To address spatial alignment issues, the receiving vehicle transforms these data levels from the sender's local coordinate system into its own. This transformation is achieved by constructing a matrix using additional information such as LiDAR sensor poses and GPS/IMU readings.

E. Comparative Analysis

Through a comprehensive review of collaboration types in CP approaches, each level offers distinct advantages and challenges. Early collaboration, while providing the richest information from various agents, demands substantial bandwidth, and methods to address time latency at the raw data level remain underexplored. In contrast, late collaboration is bandwidth-efficient but sacrifices significant scene semantic context, resulting in decreased performance and robustness against noise. Intermediate collaboration balances efficiency and accuracy, enhancing noise robustness within the system. To optimize further, hybrid collaboration allows dynamic

TABLE IX

OVERVIEW OF THE METHODS FOR COLLABORATIVE SEMANTIC SEGMENTATION (CSS). V: VEHICLE, I: INFRASTRUCTURE, UAV: UNCREWED AERIAL VEHICLE, RAW: RAW DATA FUSION, TRAD FEAT: TRADITIONAL FEATURE FUSION, ATTEN FEAT: ATTENTION FEATURE FUSION

Method	Publication	Year	Modality	Agents	**Representation**	Scheme	Fusion	Code
When2com [17]	CVPR	2020	Camera	UAV	**2D**	Intermediate	Trad Feat	✓
Who2com [91]	ICRA	2020	Camera	UAV	**2D**	Intermediate	Trad Feat	✓
MASH [92]	IROS	2021	Camera	UAV	**2D**	Intermediate	Atten Feat	✗
GenBEV [93]	ISPRS	2023	LiDAR	V	**BEV**	Early	Raw	✓
CoBEVT [94]	CoRL	2023	Camera	V	**BEV**	Intermediate	Trad Feat	✓
VICSS [95]	VTC	2023	LiDAR	V,I	**3D**	Intermediate	Atten Feat	✗
CoHFF [96]	CVPR	2024	Camera	V	**3D**	Intermediate	Atten Feat	✓

TABLE X

OVERVIEW OF THE METHODS FOR COLLABORATIVE OBJECT TRACKING (COT). V: VEHICLE, I: INFRASTRUCTURE, OBJ: OBJECT, COD: COLLABORATIVE OBJECT DETECTION, ATTEN FEAT: ATTENTION FEATURE FUSION, OBJ FUSION: OBJECT-LEVEL FUSION

Method	Publication	Year	Modality	Entity	Scheme	Shared data	**Tracker**	Fusion	Code
Track-by-det [97]	IV	2023	Agnostic	V,I	NA	Obj	**with COD**	NA	✗
HYDRO-3D [98]	T-IV	2023	LiDAR	V,I	Intermediate	Feature	**with COD**	Atten Feat	✗
FFTrack [99]	CVPR	2023	LiDAR	V,I	Intermediate	Feature	**with COD**	Atten Feat	✗
MOT-CUP [100]	RA-L	2024	Agnostic	V	NA	Obj	**with COD**	NA	✗
DMSTrack [101]	ICRA	2024	Agnostic	V	Late	Obj	**without COD**	Obj Fusion	✓

combinations of early, intermediate, or late collaboration based on accuracy demands. However, implementing hybrid frameworks is complex, mainly due to the challenges of managing heterogeneous data sources.

VI. PERCEPTION TASKS (RQ1)

There are various critical perception tasks that can benefit from a collaborative approach, including object detection, object tracking, motion prediction, semantic segmentation, and lane detection. Object Detection (OD) identifies and locates objects within a sensor frame, establishing a foundation for further perception processes. Object Tracking (OT) involves monitoring the dynamic status of an object across multiple frames, while Motion Prediction (MP) aims to forecast the future movements or intentions of an object. Semantic Segmentation (SS) plays a crucial role in scene understanding, helping CAVs identify drivable areas and provide essential information for subsequent tasks. Lane Detection (LD) is integral to determining road boundaries and lane markings, enabling CAVs to comprehend the geometry of the road network. This section provides a comprehensive overview of collaborative methods used in detection, tracking, motion prediction, semantic segmentation, and lane detection. Additionally, it introduces concepts of Multi-Task (MT) and Task-agnostic (TA) pipelines, which are pivotal in enhancing the efficiency and accuracy of vehicle perception systems.

Furthermore, the subcategories for each perception task are provided, with classifications based on representation formats. For example, OD and SS are categorized into 2D, 3D, and BEV representations. MT is divided into trajectory and BEV map representations, while LD is classified into curve-model and BEV map representations. OT is further categorized into tracking with Collaborative Object Detection (COD) and tracking without COD.

A. Collaborative Object Detection

Object detection is a fundamental perception task that focuses on identifying and locating relevant objects from raw sensor data. Typically, object detection results are presented as bounding boxes, each labeled with the corresponding object category. These bounding boxes can vary in representation: they may appear in 2D, BEV, or 3D formats. 2D bounding boxes, often used in camera-based 2D object detection, capture object on image plane. BEV representation disregards height and emphasizes the spatial layout of dynamic objects on the road plane, which is often sufficient for downstream tasks such as planning. The 3D format includes height and z-axis position, offering a more comprehensive view of the scene. This section discusses COD across these different representations, with 3D being the most prevalent form in COD applications. All papers on COD that meet our criteria are summarized in Tables XXIX.

1) 2D: Collaborative 2D object detection focuses on recognizing individual objects across multiple viewpoints on the image plane, which is particularly challenging. For instance, Khalifa et al. [88] propose a multi-view pedestrian detection approach that proceeds through a sequence of steps: monocular detection, geometric transformation, detection matching, and detection fusion. Similarly, Marez et al. [89] introduce a general COD framework, CP Faster-RCNN, designed to detect both vehicles and pedestrians. This framework extracts features from multiple viewpoints and uses an alignment module to warp them, followed by feature fusion to generate detection results. Mao et al. [90] present MoRFF, a multi-view object detection pipeline that reduces communication costs by matching deep features rather than image data.

2) BEV and 3D: BEV and 3D bounding boxes are widely used to represent dynamic objects in autonomous driving applications. The BEV representation simplifies the 3D

TABLE XI

OVERVIEW OF THE METHOD FOR COLLABORATIVE MOTION PREDICTION (CMP). V: VEHICLE, I: INFRASTRUCTURE, RAW: RAW DATA FUSION, TRAD FEAT: TRADITIONAL FEATURE FUSION, ATTEN FEAT: ATTENTION FEATURE FUSION, OBJ FUSION: OBJECT-LEVEL FUSION, GRAPH: GRAPH-BASED FUSION

Method	Publication	Year	Modality	Entity	Scheme	**Representation**	Fusion	Code
V2VNet [84]	ECCV	2020	LiDAR	V	Intermediate	**Trajectory**	Graph	✓
V2VNet-Robust [18]	CoRL	2021	LiDAR	V	Intermediate	**Trajectory**	Hybrid(Atten Feat, Graph)	✗
Late-early [60]	IEEE T-ITS	2024	LiDAR	V,I	Hybrid	**Trajectory**	Hybrid(Raw,Obj)	✓
BEV-V2X [102]	IEEE T-IV	2023	Camera	V,I	Intermediate	**BEV Map**	Atten Feat	✗
V2XFormer [103]	AAAI	2024	Camera	V,I	Intermediate	**BEV Map**	Trad Feat	✓

TABLE XII

OVERVIEW OF THE METHODS FOR CLD. V: VEHICLE, TRAD FEAT: TRADITIONAL FEATURE FUSION

Method	Publication	Year	Modality	Entity	Scheme	**Representation**	Fusion	Code
Co-mapping [107]	IEEE CAVS	2020	Camera	V	Late	**Curve-model**	Kalman filter	✗
CoLD Fusion [108]	IEEE IV	2023	Agnostic	V	Late	**Curve-model**	Spline-based Fusion	✗
LaCPF [109]	ROBOT AUTON SYST	2024	Agnostic	V	Late	**BEV map**	Trad Feat	✗

TABLE XIII

OVERVIEW OF METHODS FOR MULTI-TASK PIPELINE AND TASK-AGNOSTIC PIPELINE. OD: OBJECT DETECTION, OT: OBJECT TRACKING, MP: MOTION PREDICTION, AP: ACCIDENT PREDICTION, SS: SEMANTIC SEGMENTATION, V: VEHICLE, I: INFRASTRUCTURE, RAW: RAW DATA FUSION, TRAD FEAT: TRADITIONAL FEATURE FUSION, ATTEN FEAT: ATTENTION FEATURE FUSION, OBJ FUSION: OBJECT-LEVEL FUSION, GRAPH: GRAPH-BASED FUSION

Method	Publication	Year	Modality	Entity	Scheme	Fusion	Task	Code
V2VNet [84]	ECCV	2020	LiDAR	V	Intermediate	Graph	OD,MP	✗
Robust V2VNet [18]	CoRL	2021	LiDAR	V	Intermediate	Atten Feat, Graph	OD,MP	✗
BEV-V2X [102]	IEEE T-IV	2023	Agnostic	V, I	Intermediate	Atten Feat	SS,MP	✗
HYDRO-3D [98]	IEEE T-IV	2023	LiDAR	V	Intermediate	Atten Feat	OD,OT	✗
FF-Tracking [99]	CVPR	2023	LiDAR, Camera	V, I	Intermediate	Trad Feat	OD,OT	✓
CoBEVT [94]	CoRL	2023	Camera	V	Intermediate	Trad Feat	OD,SS	✓
V2XFormer [103]	AAAI	2024	LiDAR, Camera	V, I	Intermediate	Trad Feat	OD,MP,AP	✓
Late-early [60]	IEEE T-ITS	2024	Camera	V, I	Hybrid	Hybrid(Raw,Obj)	OD,MP	✓
STAR [110]	CoRL	2022	LiDAR	V	Intermediate	Trad Feat	Task-agnostic	✓
Core [111]	ICCV	2023	LiDAR	V	Intermediate	Trad Feat	Task-agnostic	✓

TABLE XIV

OVERVIEW OF THE METHODS FOR ADDRESSING POSE ERROR. V: VEHICLE, I: INFRASTRUCTURE, RAW: RAW SENSOR DATA, FEAT: FEATURE, OBJ: OBJECT-LEVEL DATA

Method	Publication	Year	Modality	Entity	**Data**	Pose correction approach	Code
JointPerception [8]	IEEE Sensors	2022	LiDAR	V	**Raw**	ICP Point cloud registration	✗
FastClustering [81]	Cogn. Comput.	2024	LiDAR	V	**Raw**	ICP Point cloud registration	✗
Robust V2VNet [18]	CoRL	2021	LiDAR	V	**Feat**	Markov random field	✗
BEV-V2X [102]	IEEE T-IV	2023	Agnostic	V,I	**Feat**	Global spatial aware attention	✗
FeaCo [51]	ACM MM	2023	LiDAR	V	**Feat**	Proposal Centers Matching	✓
MoRFF [90]	IEEE VTC	2023	Camera	V	**Feat**	Multi-view feature matching	✗
FPV-RCNN [87]	IEEE RA-L	2024	LiDAR	V	**Feat**	Semantic keypoint feature matching	✓
Co-perception [15]	IEEE IV	2023	Agnostic	V	**Obj**	optimal transport theory	✗
CoAlign [55]	ICRA	2023	LiDAR	V	**Obj**	Agent-Object Pose Graph Optimization	✓
FreeAlign [61]	ICRA	2024	LiDAR	V	**Obj**	Graph matching	✓

bounding box by disregarding the height dimension, making it especially useful in camera-based pipelines that utilize BEV features for detection. For instance, Hu et al. [69] present CoCa3D, a camera-only CP pipeline that extracts BEV features through a depth estimation module and voxel transformation module, subsequently decoding these features to predict object locations. Similarly, LiDAR-based pipelines can also leverage BEV features by collapsing 3D voxel feature into a BEV format, which avoids computationally demanding 3D convolutions. Wei et al. [48] introduce CoBEVFlow which utilze BEV features to predict detection result as well as predict the flow of BEV boxes. BEV features are also advantageous in LiDAR-camera pipelines, as they facilitate the alignment and fusion of multi-modal data. For

TABLE XV
OVERVIEW OF METHODS FOR ADDRESSING LATENCY AT THE FEATURE LEVEL. V: VEHICLE, I: INFRASTRUCTURE

Method	Publication	Year	Modality	Entity	Approach	Code
SyncNet [19]	ECCV	2022	LiDAR	V	Time-series prediction	✓
UMC [114]	ICCV	2023	LiDAR	V	Time-series prediction	✓
CoBEVFlow [48]	NeurIPS	2023	LiDAR	V	BEV/ROI flow prediction	✓
FFNet [115]	NeurIPS	2023	LiDAR	V,I	Feature flow prediction	✓
How2comm [52]	NeurIPS	2023	LiDAR	V	Feature flow prediction	✓
FF-Tracking [99]	CVPR	2023	LiDAR, Camera	V,I	Feature flow prediction	✓
V2X-INCOP [20]	IEEE T-IV	2024	LiDAR	V,I	Feature flow prediction	✗

TABLE XVI
OVERVIEW OF METHODS FOR ADDRESSING COMMUNICATION EFFICIENCY. V: VEHICLE, I: INFRASTRUCTURE

Method	Publication	Year	Modality	Entity	Approach for comm. efficiency	Code
F-Transformer [47]	IEEE SEC	2019	LiDAR	V	Data selection	✗
MASH [92]	IROS	2021	Camera	UAV	Data selection	✗
Where2comm [75]	NeurIPS	2022	LiDAR, Camera	V	Data selection, Cooperator selection	✓
CoCa3D [69]	CVPR	2023	Camera	V,I	Data selection	✓
DFS [37]	IEEE ITSC	2023	LiDAR	V,I	Data selection	✗
What2comm [53]	ACM MM	2023	LiDAR	V,I	Data selection	✗
How2comm [52]	NeurIPS	2023	LiDAR	V	Data selection, Data compression	✓
FPV-RCNN [87]	IEEE RA-L	2024	LiDAR	V	Data selection	✓
EdgeCooper [10]	IEEE JSAC	2024	LiDAR	V,I	Data selection	✗
SemanticComm [119]	J. Franklin Inst.	2024	LiDAR	V	Data selection	✗
PillarAttention [68]	IEEE IoT-J	2024	LiDAR	V,I	Data selection	✗
CenterCoop [59]	IEEE RA-L	2024	LiDAR	V,I	Data selection	✗
AFS-COD [11]	IEEE CAVS	2020	LiDAR	V	Data compression	✗
Slim-FCP [33]	IEEE IoT-J	2022	LiDAR	V	Data compression	✗
When2com [17]	CVPR	2020	Camera	UAV	Cooperator selection	✓
Who2com [91]	ICRA	2020	Camera	UAV	Cooperator selection	✓
Co3D [50]	IEEE T-ITS	2023	LiDAR	V,I	Cooperator selection	✗

TABLE XVII
OVERVIEW OF METHODS FOR ADDRESSING DOMAIN SHIFT. V: VEHICLE, I: INFRASTRUCTURE

Method	Publication	Year	Modality	Entity	Domain gap	Approach for bridging gap	Code
FDA [120]	ICRA	2024	LiDAR	V,I	Dataset domain	Learnable Feature Compensation	✗
DI-V2X [21]	AAAI	2023	LiDAR	V,I	LiDAR sensor domain	Domain invariant distillation framework	✓
HPL-ViT [58]	ICRA	2024	LiDAR	V	LiDAR sensor domain	Heterogeneous Graph-attention	✗
DUSA [57]	ACM MM	2023	LiDAR	V,I	Sim2Real domain	Sim/Real-invariant features	✓
S2R-ViT [62]	ICRA	2024	LiDAR	V	Sim2Real domain	Domain invariant feature learning	✗

TABLE XVIII
OVERVIEW OF METHODS FOR ADDRESSING THE PROBLEM OF HETEROGENEITY. V: VEHICLE, I: INFRASTRUCTURE

Method	Publication	Year	Modality	Entity	Heterogeneity	Approach for addressing Hetero.	Code
MPDA [121]	ICRA	2023	LiDAR	V,I	Model	Cross-Domain Transformer: unify the feature patterns from different agents	✓
HGAN [80]	IEEE PAAP	2022	LiDAR, Camera	V,I	Modality	Data format alignment: virtual 3D points from RGB	✗
HM-ViT [22]	ICCV	2023	LiDAR, Camera	V	Modality	Feature interaction: 3D Graph Attention	✓
HEAL [77]	ICRA	2024	Agnostic	V	Modality	Feature alignment: Backward alignment mechanism	✓

example, Yin et al. [74] present V2VFormer++, a multi-modal detection pipeline that first fuses BEV features from LiDAR and camera data at the entity level and then combines the multi-modal features across entities in the CP fusion step, resulting in a streamlined and unified fusion process in BEV space. In addition to BEV, 3D bounding boxes are widely used in LiDAR-only pipelines and occasionally in camera-only approaches. For instance, Wang et al. [71] introduce EMIFF, a camera-based pipeline that directly employs 3D voxel features to estimate the 3D position and dimensions of objects.

TABLE XIX

OVERVIEW OF ALL PUBLICLY AVAILABLE REAL WORLD DATASETS FOR CP THAT INCLUDE INFRASTRUCTURE PERSPECTIVE. L INDICATES LIDAR AND RGB DENOTES CAMERA SENSOR IN THE MODALITIES. FOR THE TASKS THE DATASETS INCLUDE OBJECT DETECTION (OD) – 3D IF NOT INDICATED OTHERWISE –, OBJECT TRACKING (OT), MOTION PREDICTION (MP) AND DOMAIN ADAPTION (DA)

Dataset	Year	Collaboration	Modalities	Task	Location	# classes	# co-frames	Diversity
T&J [1]	2019	V2V	L	OD	USA	NA	100	Not described
I2V-MVPD [88]	2020	V2I	RGB	OD (2D)	Tunisia	1	4.7k	Weather
DAIR-V2X-C [124]	2022	V2I	L & RGB	OD	China	10	13k	Weather, daytime
V2V4Real [127]	2023	V2V	L & RGB	OD, OT, DA	USA	5	10k	Scenario
DAIR-V2X-Seq [99]	2023	V2I	L & RGB	OD, OT, MP	China	10	7.5k	Daytime
LUCOOP [126]	2023	V2V	L	OD, MP	Germany	4	13.5k	Weather, daytime
HoloVIC [125]	2024	V2I	L & RGB	OD, OT	China	3	100k	Scenarios
TUMTrafV2X [128]	2024	V2I	L & RGB	OD, OT	Germany	8	1k	Daytime

TABLE XX

ADDITIONAL TECHNICAL INFORMATION FOR THE PUBLICLY AVAILABLE REAL WORLD DATASETS. A "-" INDICATES THE ABSENCE OF INFORMATION IN THE REFERENCED PUBLICATION

Dataset	CAVs	Localization	Synchronisation
T&J [1]	2	NA	NA
I2V-MVPD [88]	1	GPS	async (∼30ms)
DAIR-V2X-C [124]	1	Hybrid	async (10∼30ms) + sync (<10ms)
V2V4Real [127]	2	Hybrid	async (<50ms)
DAIR-V2X-Seq [99]	1	Hybrid	async + sync (<10ms)
LUCOOP [126]	3	Hybrid	NA
HoloVIC [125]	1	RTK/INS	sync (<10ms)
TUMTrafV2X [128]	1	RTK/INS	NA

B. Collaborative Semantic Segmentation

Semantic segmentation is a process designed to assign a semantic class label to every pixel in an image or every point in a LiDAR scan. This technique offers a granular understanding of scenes, going beyond object detection that typically uses bounding boxes to localize objects. Semantic segmentation facilitates the precise delineation of object boundaries and enables the identification of multiple instances within the same scene. However, visual occlusions can create areas where semantic labels cannot be accurately predicted. Through V2X collaboration, CAVs can extend their FOV and supplement the semantic labels of occluded areas, thus achieving a more comprehensive understanding of their surroundings. This section summarizes and categorizes Collaborative Semantic Segmentation (CSS) approaches based on their representation format. All papers on CSS that meet the selection criteria are listed in Table IX.

1) 2D: 2D semantic segmentation directly labels pixels within the 2D image plane. For instance, Liu et al. [91] introduce the Who2com framework, a pioneering collaborative approach to 2D semantic segmentation. This framework utilizes observations from multiple agents, including RGB images, aligned dense depth maps, and poses, to produce a 2D semantic segmentation mask for each agent. Additionally, Liu's subsequent When2com approach achieves improved performance with reduced bandwidth requirements [17]. In 2021, Glaser et al. [92] introduce a novel pipeline that operates solely on raw image data, showing superior performance particularly in scenarios with image occlusions. This method employs an attention mechanism to identify visually similar patches across different perspectives, a crucial step when depth and pose information are absent.

2) BEV: BEV semantic segmentaiton involves creating the top-down semantic map of the environment around a vehicle. In 2023, Yuan et al. [93] present GenBEV, the first BEV collaborative segmentation approach based on LiDAR. In this model, 3D voxel features, extracted by a backbone network, are projected onto a BEV map and processed by a task-specific head to segment both static road elements and dynamic objects. For camera-based BEV segmentation, 2D image feature is typically converted into a top-down perspective by depth estimation. For instance, Xu et al. [94] present the CoBEVT, a framework that enables collaborative generation of BEV map predictions. CAVs extract BEV features using the SinBEVT module and shares them with others. Received features are transformed to match the receiving vehicle's coordinate system using the FuseBEVT module, which integrates fused axial attention (FAX) to efficiently manage local-global interactions. Local attention resolves pixel correspondence on occluded objects, while global attention assimilates contextual information such as road topology and traffic density.

3) 3D: 3D semantic segmentation provides a more detailed understanding of the environment by incorporating not only road-plane information but also the height and spatial dimensions of objects. For instance, Liu et al. [95] introduce the first vehicle-infrastructure CSS framework. This innovative approach begins by transforming the point cloud data from infrastructure sensors into the vehicle's coordinate system, followed by a feature extraction process. The extracted features are then compressed and transmitted to the vehicle. Upon reception, these features are divided into two subsets based on whether they fall inside or outside the overlapping FOV. Each subset is processed separately to extract valuable information, then recombined and concatenated with the vehicle's own data. The integrated vehicle-infrastructure features are subsequently fed into a Multilayer Perceptron (MLP) to predict the class labels of the points. Experiments conducted on a synthetic dataset demonstrate that the framework outperforms several classical single-vehicle LiDAR semantic segmentation algorithms, showcasing its enhanced performance and utility. Besides LiDAR, RGB cameras also support 3D semantic segmentation by labeling occupied voxels semantically. Song et al. [96] present CoHFF framework, the first to explore

TABLE XXI

GENERAL OVERVIEW OF SYNTHETIC DATASETS FOR CP. L INDICATES LIDAR AS MODALITY. THE DATASETS SUPPORT OBJECT DETECTION (OD), OBJECT TRACKING (OT), SEMANTIC SEGMENTATION (SS), MOTION PREDICTION (MP) AND ACCIDENT PREDICTION (AP)

Dataset	Year	Collaboration	Modalities	Task	# classes	# co-frames	Source	Diversity
CODD [44]	2021	V2V	L	OD	2	13k	CARLA	Scenario
V2X-Sim1.0 [45]	2021	V2V	L & RGBD	OD	3	10k	CARLA & Sumo	Scenario
OPV2V [130]	2022	V2V	L & RGB	OD	1	12k	OpenCDA	Scenario
V2X-Sim 2.0 [131]	2022	V2V & V2I	L & RGBD	OD, OT, SS	3	10k	CARLA & Sumo	Scenario
V2XSet [16]	2022	V2V & V2I	L	OD	1	12k	OpenCDA	Scenario
IRV2V [48]	2023	V2V	L & RGB	OD	NA	8k	CARLA	Scenario
OPV2V-H [77]	2024	V2V	L & RGBD	OD	1	10k	OpenCDA	Scenario
Semantic OPV2V [96]	2024	V2V	L & RGBD	SS	12	10k	OpenCDA	Scenario
DeepAccident [103]	2024	V2V & V2I	L & RGB	OD, OT, SS, MP, AP	2	57k	CARLA	Weather, daytime

TABLE XXII

OVERVIEW OF MORE TECHNICAL ASPECTS OF THE SYNTHETIC DATASETS. DEFAULT SYNCHRONIZATION MEANS PERFECT SYNCHRONIZATION

Dataset	# CAVs	Synchronization	GNSS/IMU
CODD [44]	2	Default	No
V2X-Sim1.0 [45]	2-5	Default	Yes
OPV2V [130]	2-7	Default	Yes
V2X-Sim 2.0 [131]	2-5	Default	Yes
V2XSet [16]	2-7	Default	Yes
IRV2V [48]	2-5	Asynchronous	Yes
OPV2V-H [77]	2-7	Default	Yes
Semantic OPV2V [96]	2-7	Default	Yes
DeepAccident [103]	4	Default	No

TABLE XXIII

OVERVIEW OF SIMULATION EXPERIMENTS. SUPPORTED TASKS ARE OBJECT DETECTION (OD), OBJECT TRACKING (OT), SEMANTIC SEGMENTATION (SS), MOTION PREDICTION (MP), LANE DETECTION (LD), MAP FUSION (MF). A "-" INDICATES THE ABSENCE OF INFORMATION IN THE REFERENCED PUBLICATION

Reference	Year	V2X	Simulator	Task
[11]	2020	V2V	Volony (CARLA based)	OD
[13]	2021	V2V	CARLA	MF
[92]	2021	V2V (drones)	AirSim	SS
[14]	2022	V2V	Based on other dataset	OT
[8]	2022	V2V	CARLA	OD
[65]	2022	V2V	Gazebo	OD
[37]	2023	V2V & V2I	CARTI	OD
[95]	2023	V2I	CARLA	SS
[102]	2023	V2V & V2I	NA	MP
[108]	2023	V2V	CARLA & Resist	LD
[38]	2023	V2V	CARLA & Resist	OD
[76]	2023	V2V	CARLA	OD
[58]	2023	V2V	OpenCDA	OD
[132]	2023	V2V & V2I	CARLA	OD
[10]	2024	V2V & V2I	CARLA, Sumo, NS3	OD
[81]	2024	V2V	OpenCDA	OD

collaborative semantic occupancy prediction. It consists of four modules: occupancy prediction, semantic segmentation, V2X feature fusion, and task feature fusion. Initial RGB data is processed for depth estimation and then transformed into a voxel representation, supplemented by a 3D occupancy encoder. The semantic segmentation task net maps RGB-derived 2D semantic features onto the 3D space using deformable cross-attention. These features are projected onto orthogonal planes, optimizing bandwidth usage. V2X feature fusion updates these features with input from various agents, enhancing the perception beyond the ego vehicle's observations. The task-fusion module combines multi-agent features to reconstruct a comprehensive semantic occupancy grid, effectively mitigating issues caused by visual occlusion.

C. Collaborative Object Tracking

Object tracking involves locating and following object trajectories across sequences of video frames or point cloud data. Accurate tracking enables determination of an object's position, velocity, and acceleration, collectively understood as its motion status. Challenges in object tracking, such as dynamic changes in appearance, occlusions, and complex motion patterns, necessitate robust algorithms for continuous and precise tracking. Multi-view collaboration is a promising solution to address occlusions and maintain continuous tracking. This section categorizes collaborative object tracking into two approaches: tracking with Collaborative Object Detection (COD) and tracking without COD. Tracking with COD integrates closely with collaborative detection outcomes, enhancing subsequent perception processes. Alternatively, tracking without COD offers flexibility by fusing perception results from multiple agents independently. Both methods predominantly utilize Kalman filters and their variants for tracking and incorporate uncertainty propagation to refine their tracking processes. All papers on Collaborative Object Tracking (COT) that meet the selection criteria are summarized in Table X.

1) Tracking With COD: Tracking with COD involves performing tracking based on results from collaborative object detection. For instance, Su et al. [97] propose a 3D multi-object tracking (3D-MOT) framework that utilizes results from collaborative detection. The process begins with the tracker receiving collaborative detection results, followed by the estimation of object states at the next frame using a Kalman filter. The states are then matched to update the tracked object's status and initialize any new objects detected. This approach significantly reduces false negatives and positives compared to individual 3D-MOT setups. Additionally, Su et al. [100] introduce a method to address uncertainty in detection, termed MOT-CUP. This framework quantifies uncertainty using conformal prediction, assuming a Gaussian distribution, which is incorporated into a Standard Deviation-based Kalman Filter (SDKF) for enhanced prediction accuracy.

TABLE XXIV
OVERVIEW OF THE EVALUATION ENVIRONMENTS FOR REAL-WORLD AND SIMULATION SCENARIOS

Scenario		Urban	Rural	Highway
Real-world	V2V	[126], [127], [1]	NA	NA
	V2I	[88], [124], [125], [128], [99], [66], [35]	NA	NA
	Both	NA	NA	NA
Simulation	V2V	[130], [77], [96], [45], [48], [44]	[130], [77], [96], [45], [44]	[130], [77], [96], [45]
	V2I	NA	NA	NA
	Both	[103], [131], [16]	[131], [16]	[131], [16]

TABLE XXV
OVERVIEW OF ROAD ENVIRONMENT FOR EVALUATION

Road Environment	References
Roundabouts	[88]
Straight roads with curves	[66], [88], [99], [125]–[127]
Cross intersection	[35], [66], [99], [124], [125], [128]
T-Junction	[99], [126]
Parking lots	[1], [35]

TABLE XXVI
OVERVIEW OF GENERAL EVALUATION METRICS FOR COLLABORATIVE PERCEPTION TASKS

Task	Metrics
Object Detection	**mean Average Precision (mAP)**, Average Precision (AP), Recall, Precision, Average Recall (AR), mean Average Orientation Similarity (mAOS)
Object Tracking	**AMOTA**, **AMOTP**, sAMOTA, MOTA, MOTP, HOTA, Recall, FP, FN, ID F1 Score, Mostly Tracked Trajectories (MT), Mostly Lost Trajectories (ML), Negative Log Likelihood (NLL), Continuous Ranked Probability Score (CRPS)
Prediction	**minADE**, **minFDE**, Video Panoptic Quality (VPQ), Accident Prediction Accuracy(APA), L2 Displacement Error, End-Point Error (EPE), strict/ relaxed accuracy (AccS/ AccR), outlier ratio (ROutliers), and Missing Rate (MR)
Semantic Segmentation	**Mean Intersection over Union (mIoU)**, Intersection over Union (IoU)
Lane Detection	**Mean Squared Error MSE**, Maximum Error, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Intersection over Union (IoU)

2) Tracking Without COD: Tracking without COD utilizes lists of detected objects from multiple agents to enable cooperative tracking. For instance, Chiu et al. [101] present DMSTrack framework, a differentiable multi-sensor Kalman filter facilitates 3D multi-object tracking. Uniquely, this framework decentralizes the prediction of object state covariances, allowing each vehicle to independently predict uncertainties associated with its detections. These detected object states, along with their predicted uncertainties, are then transformed from local to global coordinate systems before being shared with neighboring vehicles. Once integrated, these data inform the Kalman filter's prediction and update stages, allowing for continuous and robust tracking by effectively managing the detection uncertainties from various agents.

D. Collaborative Motion Prediction

Motion prediction involves forecasting the future states of moving objects using historical data, a critical capability for autonomous navigation. Accurate predictions of dynamic entities' trajectories allow systems to make right decisions, thereby enhancing safety and operational efficiency. The task becomes increasingly complex in environments with multiple interacting agents due to the nonlinear and unpredictable nature of agent interactions.

Collaborative motion prediction leverages the collective intelligence of multiple observing agents, integrating diverse data sources to enhance the accuracy and robustness of predictions. This cooperative approach not only mitigates the effects of individual sensor occlusions but also provides a more reliable prediction framework compared to isolated mechanisms. Motion prediction can be descirbed as forecasting the trajectory of bounding boxes or forcasting the BEV map. All papers on Collaborative Motion Prediction (CMP) that meet the selection criteria are summarized in Table XI.

1) Trajectory: Dynamic objects in environment can be represented through the bounding boxes with attritubes such as position and shape. In this case, motion prediction means to predict a sequence of future position of the bounding boxes, known as the trajectory. For instance, Wang et al. [84] introduce V2VNet, a pioneering collaborative framework designed for simultaneous perception and prediction, termed Perception and Prediction (P&P). This approach not only enhances performance but also increases computational efficiency compared to traditional two-step processes. V2VNet extends individual P&P capabilities by integrating V2V communication. The model captures multi-scale historical data using Inception-like convolutional blocks [104] for accurate forecasting. After integrating data across different agents, the combined feature map is processed through dual networks that deliver detection and motion forecasting outcomes. In 2021, Vadivelu et al. [18] enhance V2VNet by addressing pose errors, thus improving accuracy. Additionally, Dao et al. [60] present a LiDAR-based method for scene flow prediction, called Aligner, which can be adapted for motion forecasting. Aligner predicts the movement of point-wise features extracted from LiDAR point clouds, achieving precise scene flow predictions.

2) BEV Map: BEV map can naturally combine the static road map and dynamic object map together, which benefits

TABLE XXVII
OVERVIEW OF CUSTOM EVALUATION METRICS FOR COLLABORATIVE PERCEPTION TASKS

Aspect	Metrics	Description
Communication	Average message size	Measures the communication cost in units such as Byte, KB, MB, or Mbps.
Communication and perception improvement	BIS	Bandwidth Improvement Score (BIS): A ratio of relative improvement in overall accuracy over bandwidth usage. Smaller bandwidth usage and larger improvements in overall accuracy lead to higher scores.
	AIB	Average-precision-improvement-to-bandwidth-usage (AIB): evaluate the trade-off between detection performance improvement and bandwidth usage of the proposed framework.
	RB Ratio	Recall/Bandwidth (RB) Ratio: Evaluates bandwidth efficiency relative to recall performance.
Perception improvement	Marginal Gain	Measures the the performance increase when an additional agent joins the collaboration.
Perception performance of invisible agents	ARSV	Average Recall of agents visible from Single-Vehicle View.
	ARCV	Average Recall of agents invisible from Single-Vehicle View but visible from Collaborative-View.
	ARCI	Average Recall of Completely-Invisible agents.
	ARTC	Average Recall of agents visible previously but occluded at present

TABLE XXVIII
ABLATION STUDIES GROUPED BY ASPECT

Aspect	Ablation studies
Communication	Bandwidth, Latency, Packet drop rate, Communication noise (SNR in dB), Probability of interruption, Compression rate
Localization	Pose error, Position error, Heading error
Scale of system	Number of CAVs, CAV rate, Ratio of Lidar/camera-equipped agents
Visual occlusion	Occlusion level
Adversarial attack	Attack ratio
Others	Object distance, Traffic density, Object speed, Ego vehicle velocity and acceleration, Number of cameras dropped

the motion prediction of objects. Wang et al. [103] introduce a camera-based framework, V2XFormer, which builds upon the capabilities of BEVerse [105]. V2XFormer utilizes the Swin-Transformer [106] to extract BEV features and incorporates a multi-task head that simultaneously addresses detection and motion prediction tasks. This model introduces the V2XFusion module, which integrates BEV features from multiple vehicles, enhancing collaborative perception capabilities. Chang et al. [102] introduce BEV-V2X, a pioneering framework for cooperative prediction of BEV occupancy grid maps. This framework represents dynamic objects and road structures within the BEV occupancy grid on the map, capturing the dynamics of the scene over time. BEV-V2X leverages historical and current BEV map data to forecast future BEV maps within a three-second timeframe.

E. Collaborative Lane Detection

Lane detection is a critical component for Advanced Driver Assistance Systems (ADAS) and automated driving (AD), as it provides essential information for path planning and vehicle control. High-definition Map (HD Map), though effective, is expensive to create, maintain, and scale. This makes real-time lane detection and online HD Map learning increasingly important. However, lane detection, like other perception tasks, faces challenges such as visual occlusion and limited perception range, particularly in urban intersections with dense traffic, where multi-agent collaboration offers a potential solution. This section categorizes collaborative lane detection into two main approaches: curve-model-based methods and BEV-map-based methods. Lane information can be represented using curve models, which are more data-efficient and require less bandwidth, or BEV segmentation, which provides pixel-level detail with higher resolution and greater robustness to noise. While both approaches offer substantial potential, they remain under-explored and require deeper investigation. All papers on Collaborative Lane Detection (CLD) that meet the selection criteria are summarized in Table XII.

1) Curve-Model-Based: Curve-model-based methods represent the lane information as mathematical curves, enabling efficient data sharing and processing. For example, Sakr et al. [107] propose a cooperative road geometry estimation framework, where sensor-rich vehicles share perceived road information with other vehicles. The road is divided into multiple connected segments, with each segment described by a clothoid-based model that uses parameters such as position, initial curvature, and curvature change rate. These parameters can be transmitted via V2X communication to extend the perception range. However, this approach does not account for fusing local lane detection data. To address this, Gamerdinger et al. [108] introduce convoy fusion and spline fusion methods, which handle scenarios with and without overlapping lanes, respectively. Convoy fusion uses a weighted mean to merge lane data, assuming that closer lane detection is more accurate. For non-overlapping segments, spline fusion reconstructs the road between visible segments to provide a complete lane model.

2) BEV-Map-Based: While these methods represent the road using segmented curves, Jahn et al. [109] propose LaCPF,

TABLE XXIX

OVERVIEW OF THE METHODS FOR COLLABORATIVE OBJECT DETECTION (COD) BASED ON BEV AND 3D REPRESENTATIONS. V: VEHICLE, I: INFRASTRUCTURE, RAW: RAW DATA FUSION, TRAD FEAT: TRADITIONAL FEATURE FUSION, ATTEN FEAT: ATTENTION FEATURE FUSION, OBJ FUSION: OBJECT-LEVEL FUSION, GRAPH: GRAPH-BASED FUSION

Paper	Modality	Scheme	Year	Entity	Fusion	Code
JointPerception [8]	LiDAR	Early	2022	V	Raw	✗
RAO [9]	LiDAR	Early	2023	V	Raw	✗
EdgeCooper [10]	LiDAR	Early	2024	V,I	Trad Feat	✗
FastClustering [81]	LiDAR	Early	2024	V	Raw	✗
F-cooper [1]	LiDAR	Intermediate	2019	V	Trad Feat	✓
AFS-COD [11]	LiDAR	Intermediate	2020	V	Trad Feat	✗
FS-COD [12]	LiDAR	Intermediate	2020	V	Trad Feat	✗
CoFF [32]	LiDAR	Intermediate	2021	V	Trad Feat	✗
SyncNet [19]	LiDAR	Intermediate	2022	V	Trad Feat	✓
PillarGrid [49]	LiDAR	Intermediate	2022	V,I	Trad Feat	✗
Slim-FCP [33]	LiDAR	Intermediate	2022	V	Trad Feat	✗
AdaptiveFeature [44]	LiDAR	Intermediate	2023	V	Trad Feat	✓
CoBEVFlow [48]	LiDAR	Intermediate	2023	V	Trad Feat	✓
DI-V2X [21]	LiDAR	Intermediate	2023	V,I	Trad Feat	✓
DFS [37]	LiDAR	Intermediate	2023	V,I	Trad Feat	✗
FFNet [115]	LiDAR	Intermediate	2023	V,I	Trad Feat	✓
CoAlign [55]	LiDAR	Intermediate	2023	V	Trad Feat	✓
VINet [132]	LiDAR	Intermediate	2023	V,I	Trad Feat	✗
FDA [120]	LiDAR	Intermediate	2024	V,I	Trad Feat	✗
HP3D-V2V [39]	LiDAR	Intermediate	2024	V	Trad Feat	✗
MACP [41]	LiDAR	Intermediate	2024	V	Trad Feat	✓
S2R-ViT [62]	LiDAR	Intermediate	2024	V	Trad Feat	✗
Select2Col [63]	LiDAR	Intermediate	2024	V	Trad Feat	✓
PillarAttention [68]	LiDAR	Intermediate	2024	V,I	Trad Feat	✗
DUSA [57]	LiDAR	Intermediate	2023	V,I	NA	✓
UMC [114]	LiDAR	Intermediate	2023	V	Graph	✓
HPL-ViT [58]	LiDAR	Intermediate	2024	V	Graph	✗
F-Transformer [47]	LiDAR	Intermediate	2019	V	Atten Feat	✗
CRCNet [43]	LiDAR	Intermediate	2022	V	Atten Feat	✗
V2X-ViT [16]	LiDAR	Intermediate	2022	V,I	Atten Feat	✓
MPDA [121]	LiDAR	Intermediate	2023	V,I	Atten Feat	✓
Co3D [50]	LiDAR	Intermediate	2023	V,I	Atten Feat	✗
FeaCo [51]	LiDAR	Intermediate	2023	V	Atten Feat	✓
How2comm [52]	LiDAR	Intermediate	2023	V	Atten Feat	✓
LCRN [54]	LiDAR	Intermediate	2023	V	Atten Feat	✗
SCOPE [56]	LiDAR	Intermediate	2023	V	Atten Feat	✓
What2comm [53]	LiDAR	Intermediate	2023	V,I	Atten Feat	✗
CenterCoop [59]	LiDAR	Intermediate	2024	V,I	Atten Feat	✗
V2X-INCOP [20]	LiDAR	Intermediate	2024	V,I	Atten Feat	✗
MKD-Cooper [66]	LiDAR	Intermediate	2024	V	Atten Feat	✓
Self-Adaptive [136]	LiDAR	Intermediate	2024	V	Atten Feat	✗
SemanticComm [119]	LiDAR	Intermediate	2024	V	Atten Feat	✗
V2VFormer [34]	LiDAR	Intermediate	2024	V	Atten Feat	✗
FL-Dynamic [13]	LiDAR	Late	2021	V	Obj	✓
Env-T2TF [14]	LiDAR	Late	2022	V	Obj	✗
Co-perception [15]	LiDAR	Late	2023	V	Obj	✗
Among Us [23]	LiDAR	Late	2023	V	Obj	✓
Collective PV-RCNN [38]	LiDAR	Late	2023	V	Trad Feat	✗
Late-CNN [36]	LiDAR	Late	2023	V	Obj	✗
Model-Agnostic [85]	LiDAR	Late	2023	V	Obj	✓
Double-M [137]	LiDAR	Late	2023	V,I	Obj	✓
Pillar-based CP [65]	LiDAR	Hybrid	2022	V	Hybrid (Raw, Trad Feat, Obj)	✗
ML-Cooper [35]	LiDAR	Hybrid	2022	V	Hybrid (Raw, Trad Feat, Obj)	✗
FPV-RCNN [87]	LiDAR	Hybrid	2024	V	Trad Feat	✓
Hybrid-CP [64]	LiDAR	Hybrid	2024	V	Hybrid (Atten Feat, Obj)	✗
FreeAlign [61]	LiDAR	Hybrid	2024	V	Trad Feat	✓
CoCa3D [69]	Camera	Intermediate	2023	V,I	Trad Feat	✓
ActFormer [70]	Camera	Intermediate	2024	V	Atten Feat	✓
EMIFF [71]	Camera	Intermediate	2024	V,I	Trad Feat	✓
QUEST [72]	Camera	Intermediate	2024	V,I	Trad Feat	✗

a different approach with the lightweight collaborative lane detection framework. In this method, roads are represented as static BEV maps, transforming lane detection into a BEV segmentation task. Each vehicle generates its own BEV road segmentation, which is shared via V2X with neighboring vehicles. After aligning all local BEV data into the same coordinate system, a fusion process using an encoder-decoder architecture combines the data into a comprehensive segmentation result.

TABLE XXIX
(Continued.) OVERVIEW OF THE METHODS FOR COLLABORATIVE OBJECT DETECTION (COD) BASED ON BEV AND 3D REPRESENTATIONS. V: VEHICLE, I: INFRASTRUCTURE, RAW: RAW DATA FUSION, TRAD FEAT: TRADITIONAL FEATURE FUSION, ATTEN FEAT: ATTENTION FEATURE FUSION, OBJ FUSION: OBJECT-LEVEL FUSION, GRAPH: GRAPH-BASED FUSION

Paper	Modality	Scheme	Year	Entity	Fusion	Code
ViT-FuseNet [78]	LiDAR, Camera	Early	2024	V,I	Atten Feat	✗
Multi-vehicle fusion [76]	LiDAR, Camera	Intermediate	2023	V	Trad Feat	✗
V2VFusion [79]	LiDAR, Camera	Intermediate	2023	V	Trad Feat	✗
HEAL [77]	LiDAR, Camera	Intermediate	2024	V	Trad Feat	✓
HGAN [80]	LiDAR, Camera	Intermediate	2022	V,I	Hybrid	✗
Distilled Co-Graph [45]	LiDAR, Camera	Intermediate	2021	V	Graph	✓
HM-ViT [22]	LiDAR, Camera	Intermediate	2023	V	Graph	✓
PIXOR [130]	LiDAR, Camera	Intermediate	2022	V	Atten Feat	✓
Where2comm [75]	LiDAR, Camera	Intermediate	2022	V	Atten Feat	✓
MCoT [138]	LiDAR, Camera	Intermediate	2023	V	Atten Feat	✗
PAFNet [67]	LiDAR, Camera	Intermediate	2024	V,I	Atten Feat	✗
V2VFormer++ [74]	LiDAR, Camera	Intermediate	2024	V	Atten Feat	✗
TCLF [124]	LiDAR, Camera	Late	2022	V,I	Obj	✓
VICOD [86]	LiDAR, Camera	Late	2022	V,I	Obj	✗

Lane information can be represented through either curve models or BEV segmentation. Curve models are more data-efficient and require less bandwidth, while BEV segmentation provides pixel-level detail, offering higher resolution and greater robustness to noise. Both approaches have significant potential but remain underexplored.

F. Multi-Task and Task-Agnostic

Autonomous vehicle navigation requires addressing various perception tasks, from object detection to semantic segmentation. Traditionally, these tasks are performed independently, consuming significant computational resources. To optimize resource usage and enhance performance across multiple tasks simultaneously, researchers have proposed multi-task learning pipelines to address multiple perception tasks. All papers on multi-task and task-agnostic method that meet the selection criteria are summarized in Table XIII. For example, V2XFormer [103] simultaneously performs object detection, motion prediction, and accident prediction. Similarly, CoBEVT [94] handles both object detection and semantic segmentation in parallel. V2VNet [84] is also able to conduct object detection and motion prediction at the same time.

However, multi-task learning alone cannot fully address task heterogeneity issues. To tackle this challenge, researchers have proposed task-agnostic frameworks, such as Collaborative Scene Completion (CSC), which can support various downstream perception tasks. In 2022, Li et al. [110] introduce STAR, a multi-agent scene completion framework where each agent learns to reconstruct the complete scene as viewed by all agents. STAR employs a spatial-temporal autoencoder architecture with a vision transformer (ViT) backbone to extract scene features. These features from various agents are aggregated with pose awareness and then processed by a decoder to predict the complete view. STAR demonstrates compatibility with single-agent perception models, allowing for integration without additional training. This approach significantly benefits scenarios with visual occlusion. In contrast to STAR, which conducts downstream tasks on completed scene representations, Wang et al. [111] propose CORE, a novel cooperative reconstruction framework. CORE performs downstream tasks directly on collaborative features, using reconstruction as additional guidance to develop a powerful encoder and fusion module. This approach generates informative intermediate representations that are then processed by task-specific decoders for various purposes, such as detection or segmentation. CORE has shown superior performance in both 3D object detection and BEV semantic segmentation tasks while maintaining bandwidth efficiency. In conclusion, scene completion can serve as a guideline for feature learning, benefiting various downstream tasks. It can also be combined with single-agent perception models to enhance accuracy across different perception tasks.

VII. APPROACHES TO ADDRESS REALISTIC ISSUES (RQ1)

In the initial stages of research, the focus on CP primarily focused on the collaboration process and fusion strategies under ideal conditions, often relying on unrealistic assumptions such as precise localization and ideal communication conditions. However, CP algorithms encounter numerous challenges when deployed in real-world scenarios. This section summarizes these practical issues and their corresponding solutions.

A. Localization Errors

Accurate spatial alignment is essential for effective data fusion among different agents. However, errors in localization can lead to data misalignment, significantly impacting perception accuracy. To tackle this issue, researchers focus on correcting the relative pose before alignment [8], [15], [18], [51], [55], [61], [81], [87], [90], [102], the process as shown in Figure 8. Approaches to address localization errors are summarized in Table XIV. The various approaches to address this problem can be categorized into three levels: raw-sensor, object, and feature levels. These methods are comparatively analyzed below.

1) Raw-Sensor Level: To achieve accurate relative positioning between cooperative vehicles, Ahmed et al. [8] introduce a joint perception scheme that utilizes compressed point clouds. This approach employs point-to-plane Iterative Closest Point

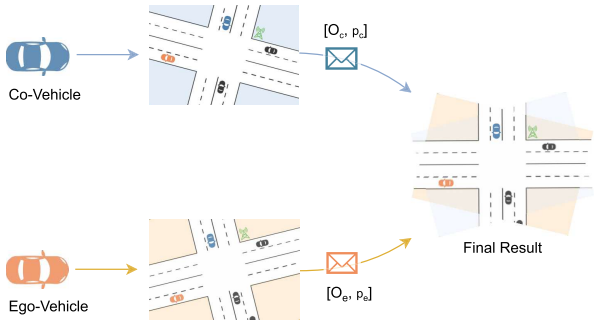

Fig. 8. Illustration of the localization error issue: The cooperative vehicle transmits its sensing data and pose to the ego vehicle. The ego vehicle corrects the relative pose before alignment using consensus derived from the sensing data. Subsequently, the sensing data from multiple agents are fused based on the corrected pose.

(ICP) registration to determine the optimal transformation matrix between the point clouds of the ego vehicle and the sender. This matrix is then used to achieve spatial alignment of the point clouds. While raw-sensor data level corrections typically provide precise pose estimations, they require the transmission of point cloud data, which consumes substantial bandwidth.

2) Object-Level: Song et al. [15] introduce the application of optimal transport theory to correct inaccurate vehicle locations and headings using only object-level bounding boxes. The pose correction process involves two stages. Given the local pose estimations of the ego vehicle and a cooperative vehicle, along with noisy measurements of perceived objects, the first step is to identify the co-visible region and associate the corresponding objects. Subsequently, an accurate transformation matrix F is estimated by optimizing the following problem:

$$\min_{\mathcal{F}} \sum_{(i,j)\in\mathcal{M}} \|x_i - \mathcal{F}(y_j)\|^2 \tag{1}$$

x represents the position vector of objects perceived by the ego vehicle (similarly, y for the cooperative vehicle), with i,j denoting associated object pairs that represent the same physical target.

Similarly, Lu et al. [55] introduce another optimization-based approach, CoAlign, commonly used in Simultaneous Localization and Mapping (SLAM) algorithms, to correct the relative pose over various timeframes once the close-loop of the pose graph is identified. CoAlign introduces an agent-object pose graph to represent the relationships between agents and objects, aiming for consistency in the object's pose from different viewpoints. This consistency is pursued by formulating and minimizing a pose consistency error optimization problem. This method not only corrects the agents' pose but also enhances the positional accuracy of perceived objects. However, pose-graph optimization depends on a good initial guess, limiting its effectiveness in the presence of large noise.

Both optimization-based methods may be constrained by an underperforming object association step, which relies on prior knowledge of the pose. To overcome this limitation, Lei et al. [61] propose a spatial alignment approach, FreeAlign, which utilizes geometric consistency of a shared object map to associate objects without prior pose knowledge. Geometric consistency implies that co-visible regions should have a similar distribution of objects, with consistent geometric characteristics between object pairs. A graph model, where nodes represent objects and edges represent relative distances, can be used to depict this relationship. By identifying the most similar graph between two agents, corresponding nodes in the two graphs represent associated objects. Subsequently, FreeAlign employs RANSAC [112] to calculate the relative pose between these object maps.

While object-level pose correction is more communication-efficient, it is generally less accurate than methods using raw-sensor data due to higher noise levels in processed object-level data.

3) Feature-Level: To balance the performance of pose correction with communication efficiency, feature-level approaches have been developed. Vadivelu et al. [18] introduce a pose regression module that estimates the relative pose through end-to-end learning, further refined by a Markov Random Field [113]. This approach has proven to enhance both object detection and motion forecasting tasks. Additionally, Chang et al. [102] develop a method that incorporates global spatially-aware attention to improve spatial alignment. This technique utilizes prior map information to compare with current BEV segmentation results, achieving more precise global positioning.

Feature matching is the most commonly used method at this level. For example, Gu et al. [51] introduce FeaCo, which utilizes a robust feature-level proposal centers matching technique to calculate an accurate transformation matrix. This matching process, inspired by ICP, minimizes the distance between original proposal centers from the ego vehicle and the transformed proposal centers from cooperative vehicles to derive the rotation matrix and translation vector. Similarly, MoRFF [90] and FPV-RCNN [87] both employ feature keypoint matching to rectify the relative pose, enhancing pose alignment. While feature matching is straightforward to implement, its accuracy is limited by the spatial resolution of the features.

Discussion: Pose correction methods vary significantly in their trade-offs among accuracy, bandwidth consumption, and robustness to noise. Approaches based on raw-sensor data typically offer the highest accuracy but require substantial bandwidth. In contrast, object-level methods are more efficient in terms of communication and computation but are less robust against noisy conditions. Feature-level approaches represent a middle ground, balancing accuracy and efficiency. The choice of a pose correction approach generally depends on the type of collaboration involved. When feature-level data is shared across agents, frameworks tend to utilize this same-level data for pose alignment, eliminating the need for external data sources. It allows for a more streamlined integration and efficient processing within the CP system.

B. Time Latency

Effective data fusion in CAVs necessitates that data be temporally aligned. In practice, Connected Autonomous Vehicle

(CAV)s synchronize their clocks with Coordinated Universal Time (UTC) primarily through Global Navigation Satellite System (GNSS) signals, while the Network Time Protocol (NTP) may serve as a fallback when GNSS is unavailable. Achieving perfect alignment, however, remains challenging due to factors such as communication delays, interruptions, heterogeneous processing times, and varying data rates across vehicles, all of which can introduce temporal misalignment. To mitigate these issues, three levels of data are utilized: object-level data, feature-level data, and occupancy-level data.

1) Object-Level: The object-level approach is frequently applied in late collaboration scenarios to adjust for the movement of dynamic objects across different time frames. This method uses motion models to predict the position of an object at the current timestamp based on its previous data frame. For example, Su et al. [97] employ the Constant Acceleration (CA) motion model to predict future positions of objects, allowing for more accurate data synchronization and integration in CAVs.

2) Feature-Level: Feature-level approaches have gained significant attention in addressing temporal alignment challenges for intermediate collaboration in CAVs. The overview of feature-level approaches to address time latency is shown in Table XV. One notable approach is the latency-aware collaborative perception system introduced by Lei et al. [19]. This system employs SyncNet, a latency compensation module utilizing Long Short-Term Memory (LSTM) networks to estimate real-time features for collaboration. SyncNet has demonstrated effectiveness in enhancing intermediate collaboration, particularly in high-latency scenarios. However, real-time feature prediction can be computationally intensive. An alternative method, proposed by Wei et al. [48], focuses on BEV flow prediction within the CP framework. This approach, called CoBEVFlow, generates spatial regions of interest (ROIs) based on received perceptual feature maps. By associating correlated ROIs across message sequences, it calculates motion vectors and estimates object positions at specific timestamps. The resulting BEV flow map is used to adjust the spatial position of features, ensuring temporal alignment with ego features for efficient aggregation. Yu et al. [115] introduce another technique called Feature Flow Net (FFNet), which employs feature flow prediction. This method describes feature changes over time, enabling direct prediction of aligned features at the current timestamp of collaborating vehicles. Similarly, the How2comm [52] framework utilizes feature flow prediction but refines it with a scale matrix. This scaling of predicted features has been shown to enhance temporal alignment effectiveness.

These feature-level approaches offer promising solutions for addressing temporal alignment challenges in CP systems. By focusing on feature prediction, BEV-ROI-flow prediction, and feature-flow prediction, researchers are developing more robust and efficient methods for CAVs to share and process perceptual information.

3) Occupancy-Level: Occupancy-Grid representation have emerged as a promising solution for 3D scene understanding, offering a more comprehensive depiction of dynamic environments compared to traditional object-level or feature-level methods. Zhang et al. [9] introduce the concept of occupancy flow prediction for temporal alignment in CP. This approach utilizes occupancy maps, which provide a more effective representation of the environment than raw point clouds and offer more detailed information compared to neural features. By predicting the flow of occupancy over time, these systems can better account for the dynamic nature of traffic scenarios and compensate for localization discrepancies between collaborating vehicles.

Discussion: Object-level approaches offer efficiency and ease of implementation, making them attractive for real-time applications. These methods typically involve sharing high-level information such as object positions and velocities. However, their effectiveness is heavily dependent on the accuracy of upstream tasks, including object tracking and motion estimation. In scenarios with significant time latency, the propagation of errors from these tasks can lead to reduced overall system performance. To mitigate such challenges, feature-level approaches offer enhanced robustness by estimating environmental features with greater temporal precision. These methods often involve sophisticated prediction algorithms that can compensate for temporal mis-alignments in data from multiple vehicles. While more complex than object-level methods, feature-level approaches offer a better balance between computational efficiency and latency mitigation. Occupancy-level approaches, particularly those employing occupancy flow prediction, deliver the most comprehensive representation of the environment by modeling dynamic occupancy states over time. These methods provide detailed environmental information and offer significant benefits for temporal alignment.

C. Communication Bandwidth Constraints

In any system requiring communication, bandwidth can become a bottleneck when multiple entities participate and actively contribute to sharing data. In CP, several entities (vehicles or infrastructure) collect, share, and aggregate perception data. In Europe, the European Telecommunications Standards Institute (ETSI) has specified functional requirements for collective awareness and cooperative perception applications. These specifications establish well-recognized networking constraints that must be considered in the design of CP algorithms. The vehicular network is ad hoc, participants establish communication in a self-organizing manner, and safety applications, such as CP, generate messages within the limits of a Packet Data Unit (PDU). The PDU size is defined by the access layer protocol, which imposes bandwidth limitations and the requirement that safety messages such as Cooperative Awareness Message (CAM) and Collective Perception Message (CPM) must fit into a single PDU, as they are broadcast only once without retransmission or forwarding.

In practice, the bandwidth available for vehicular communication is highly constrained. For instance, IEEE 802.11p/DSRC provides a theoretical data rate of up to 27 Mbps in a 10 MHz channel, but the effective throughput in dense traffic environments is typically below 10 Mbps due to protocol overhead and channel contention [116], [117]. Similarly, LTE-V2X operating in a 10 MHz channel can

achieve around 15 Mbps under ideal conditions, yet its capacity decreases substantially as the number of vehicles increases [118]. These limitations render the direct transmission of raw sensor data, such as LiDAR point clouds or high-resolution camera frames, largely infeasible in real-world deployments. This motivates the need for more efficient communication strategies that focus on transmitting intermediate features, or final perception outputs rather than raw data.

To mitigate these constraints, recent research has explored approaches such as data selection to filter transmitted information, data compression to reduce message size, and cooperator selection to restrict the number of entities participating in the CP algorithm. Table XVI provides an overview of representative methods that address communication bandwidth challenges in CP.

1) Data Selection: Data selection becomes necessary when, in the ETSI Intelligent Transportation System (ITS) scenario, there are limitations on how much data can or should be sent. Perception algorithms output data with varying accuracy, which can lead to filtering data based on the accuracy or confidence in the quality of the perception [6]. Luo et al. [10] propose a Voxelization-based strategy for LiDAR data where the detection model groups samples into voxels. When transmission is required, the number of points from a voxel is limited while still being able to represent an object. Wang et al. [47] employ data selection as a two-step process with negotiation and transmission. This method divides the view (perception field of view) into sections called pillars. Pillars with occlusion or partial occlusion require auxiliary information requested through the negotiation step. The relevant pillars are sent in the transmission step. Yang et al. [53] utilize a request-response methodology for cooperation where the ego vehicle broadcasts a request based on a filtered importance map generated from the feature map. Neighboring vehicles respond based on specificity and consistency constraints contained in the request.

2) Data Compression: Another means of reducing bandwidth is through data compression. This can be achieved by employing compression algorithms that reduce the binary representation of the same data and can be decompressed at the receiver. However, this adds overhead in both operations. Some work utilizes a change in the encoding of data, such as Slim-FCP [33], where feature maps are reduced with a negligible degradation to the recall performance of the CP algorithm. Marvasti et al. [11] utilize a Convolutional Neural Network (CNN) encoder solution to transform the feature map into a lower dimension. This compressed feature map is transmitted along with GPS information to other cooperative entities. Compression through transformation implies a trade-off between feature map accuracy (post-decompression) and transmission bandwidth requirements.

3) Cooperator Selection: Cooperator selection restricts the number of entities in the vehicular network participating in the CP algorithm. This way, the number of messages being transmitted can be reduced. However, the choice of participating entities is not trivial, since designing effective selection metrics is challenging, including determining the factors that should be incorporated into these metrics. Wang et al. [50] utilize a scoring system among participants that share feature maps encoded into query features using CNN. Each participant then scores these query features to select their communication targets. Liu et al. [17], [91] present a three-stage handshake among entities to establish a group of participants to communicate. Participants calculate a matching score based on the correlation between two entities, which represents the amount of information one entity can provide for the other.

Discussion: The three strategies - data selection, data compression, and cooperator selection - offer distinct methods for mitigating communication bandwidth constraints in CP. Data selection focuses on transmitting the most relevant information, optimizing bandwidth but risking incomplete perception if criteria are overly restrictive. Data compression achieves bandwidth efficiency through compact representations but introduces computational costs and potential loss of fidelity. Cooperator selection reduces the communication load by limiting participants, though the exclusion of key entities due to suboptimal metrics can compromise effectiveness. Combining these approaches could provide a balanced solution, leveraging their strengths to address bandwidth limitations comprehensively.

D. Communication Interruptions

Ad-hoc networks, such as the vehicular network, are prone to communication issues that lower the effectiveness of data transmission. In some cases, packets may fail to arrive due to collision, which we call communication interruption. One solution to this issue is proposed by Ren et al. [20], where missing data is estimated by prediction from a previous frame. In such cases where historical data from a known entity is available, missing frames can be estimated.

E. Domain Shifts

CP frameworks for CAVs face significant challenges due to domain shift, a problem often under-explored in the field. This section examines the approaches to address different types of domain shift caused by training data, sensor characteristics, and the transition from simulation to real-world environments (Sim2Real). An overview can be seen in Table XVII.

1) Domain Shift Caused by Training Data: Collaborative perception systems in CAVs often involve vehicles from different manufacturers, each employing its own perception pipeline. Even if these vehicles utilize the same neural network architecture for feature extraction, variations in their training data can still lead to inconsistencies in the extracted features. To address this challenge, Li et al. [120] propose the Feature Distribution-aware Aggregation (FDA) framework. The FDA framework incorporates a Learnable Feature Compensation (LFC) module, designed as an encoder-decoder architecture with skip connections, to predict and adjust residual discrepancies in the shared features. By applying this residual compensation, the shared features are enhanced before being fed into the fusion module. The FDA framework has been shown to effectively restore detection performance, even in the presence of distribution gaps, demonstrating its efficiency in maintaining reliable perception.

2) Domain Shift Caused by Sensor: CAVs from different manufacturers may be equipped with varying LiDAR sensors, which introduces inherent domain gaps in the raw sensor data. To address this issue, Li et al. [21] propose the DI-V2X framework for Vehicle-Infrastructure Collaborative 3D Object Detection. DI-V2X is designed to learn domain-invariant representations using a distillation-based approach. First, the Domain Mixing Instance Augmentation (DMA) module creates a domain-mixing 3D instance bank for both teacher and student models during training, ensuring better alignment in data representation. Following this, the Progressive Domain-Invariant Distillation (PDD) module encourages student models across different domains to progressively learn domain-invariant feature representations from the teacher model. Additionally, a Domain-Adaptive Attention Framework (DAF) is used to further close the domain gap by incorporating calibration-aware, domain-adaptive attention.

In contrast to the domain-invariant approach, Liu et al. [58] explore the use of heterogeneous graph-attention mechanisms to fuse features from different agents, each with domain-specific characteristics. In this method, vehicles equipped with different types of LiDAR are treated as heterogeneous collaborators, represented as distinct nodes in a graph. The cooperative interactions between these heterogeneous nodes are modeled as weighted edges, where the weights reflect different fusion strategies for effective collaboration across domain gaps.

3) Domain Shift Caused by Sim2Real: CP models require a large amount of labeled real-world data for training. However, collecting and annotating this data is both challenging and costly. As a result, synthetic data has gained attention due to its ease of production and cost-effectiveness. Despite these advantages, there is a significant domain gap between simulated environments and the real world, particularly in terms of appearance and content realism. This gap often leads to poor performance when models trained on simulated data are evaluated on real-world data.

To address this issue, Kong et al. [57] introduce the DUSA framework for CP. DUSA employs a Location-Adaptive Sim2Real Adapter (LSA) module to selectively aggregate features from critical locations on the feature map. It then aligns the features between simulated and real-world data using a sim/real discriminator in an adversarial training process. The aligned features are subsequently fed into the fusion module, ensuring CP remains unaffected by the Sim2Real gap. Similarly, Li et al. propose the S2R-ViT framework [62], which uses domain discriminators to extract domain-invariant features from both simulation and real-world environments. Unlike other methods, S2R-ViT not only inputs features from individual agents into the discriminator before fusion but also applies the discriminator to the fused features, enhancing feature generalization and improving model performance in real-world scenarios.

Discussion: Domain shifts can be categorized by their severity, ranging from low to high: dataset distribution, sensor characteristics, and Sim2Real discrepancies. Solutions for addressing domain shift include heterogeneous fusion, feature compensation, and domain-invariant feature learning. Heterogeneous fusion involves combining features with weights without fully eliminating the domain shift, making it less effective for larger gaps such as Sim2Real. In contrast, feature compensation and domain-invariant feature learning both aim to minimize domain gaps by generating more consistent features before fusion. Domain-invariant features can be achieved through cross-domain knowledge distillation and adversarial training, effectively bridging the gap and enhancing model performance.

F. Heterogeneity

Heterogeneity within CP systems presents a significant challenge, primarily caused by differences in sensors and perception models across agents. CAVs on the road are often manufactured by various companies, leading to differences in sensor types and data processing models across vehicles from different Original Equipment Manufacturers (OEMs). This section provides a summary of the approaches used to address both model heterogeneity and modality heterogeneity within CP systems, as listed in Table XVIII.

1) Model Heterogeneity: Current CP frameworks leverage deep neural network features to balance perception accuracy and communication bandwidth. However, these frameworks typically assume that all CAVs use identical neural networks, which is not always feasible in real-world scenarios. When features are transmitted from different models, a significant domain gap can emerge, leading to a decline in performance within CP systems.

To address this issue, Xu et al. [121] introduce the Multi-agent Perception Domain Adaptation (MPDA) framework, a plug-in module designed to work with most existing systems while preserving confidentiality. MPDA includes a learnable feature resizer to align features across multiple dimensions and a sparse cross-domain transformer for domain adaptation. A domain classifier is then used to distinguish whether the features originate from the source or target domain. Through adversarial training, the sparse cross-domain transformer learns to produce domain-invariant features. Although MPDA has shown to improve performance in heterogeneous environments, it still struggles to fully resolve significant performance drops.

2) Modality Heterogeneity: Most existing work focuses on homogeneous systems where CAVs are equipped with identical sensor types, an assumption that is unrealistic for real-world applications and significantly limits the scalability of collaboration.

To address modality heterogeneity, Zhang et al. [80] introduce the Multi-Modal Virtual-Real Fusion Transformer (MVRF) for collaborative perception. MVRF enables cross-modality cooperation between LiDAR and RGB cameras by generating virtual points from RGB images and incorporating them with LiDAR data.

In contrast to this data alignment approach, Xiang et al. [22] propose the hetero-modal Vision-Transformer (HM-ViT) for collaborative perception, which utilizes Heterogeneous 3D Graph Attention. HM-ViT separately extracts BEV features from LiDAR and camera streams, treating the features as distinct nodes on a collaborative graph. A 3D graph attention

mechanism is then applied to learn cross-modality interactions, and the updated features are fed into separate heads for final predictions from each modality.

In addition to learning cross-modality interactions, Lu et al. [77] introduce HEAL, an extensible framework for open heterogeneous CP. HEAL addresses heterogeneity by aligning features in a unified space using a multi-scale, foreground-aware Pyramid Fusion network. To integrate new agents with previously unseen models or sensor modalities, only the encoder part of the architecture on new agents needs retraining. This step aligns the new agents' BEV feature space with the unified space, offering low training costs and making the solution scalable for open heterogeneity scenarios.

Discussion: There are four primary approaches to addressing heterogeneity in CP systems. The first is to account for heterogeneity in the fusion process by learning cross-heterogeneity interactions. Another approach is to align the data format or feature space, allowing for homogeneous fusion. Additionally, rather than focusing on alignment, one can enable fusion by learning domain-invariant features across heterogeneous agents.

G. Adversarial Attacks

CP enhances scene understanding but is particularly vulnerable to adversarial attacks. Ensuring the safety of CAVs requires protecting them from such threats. While adversarial attacks have been extensively studied in the communications field, they have not been deeply explored within the context of CP frameworks.

The first study to address adversarial attacks in this domain is ROBOSAC [23], a general sampling-based framework for adversarially robust CP. ROBOSAC aims to achieve consensus among co-vehicles during collaboration, preventing significant deviations from individual perceptions. Its workflow involves several steps. First, a vehicle samples a subset of its teammates and compares the results with and without the sampled teammates. Next, it verifies the consensus across results to ensure no attackers are present. Finally, the vehicle produces a collaborative perception result. The key advantage of ROBOSAC is that it does not require prior knowledge of specific attack patterns, allowing it to be generalized to new types of adversarial attacks. ROBOSAC has been shown to significantly enhance the robustness of CP while maintaining high perception accuracy under attack.

Despite this progress, robust design against adversarial attacks in CP remains an under-explored area, requiring further investigation in the future.

VIII. Evaluation Methods for Collaborative Perception (RQ2-4)

Evaluation methods are a critical aspect of research on CP, complementing the development of CP approaches. This section provides an overview of the current evaluation methodologies employed in the surveyed studies. Section VIII-A describes the evaluation methodologies in detail, while Section VIII-B focuses on the evaluation scenarios. Additionally, Section VIII-C presents the metrics and ablation studies used to assess CP approaches.

A. Evaluation Methodology

To evaluate new algorithms for CP, various methodologies are employed. This section provides an overview of the approaches used in the surveyed papers. Real-world datasets and synthetic datasets are discussed in Sections VIII-A1 and VIII-A2, respectively. Additionally, real-world experiments and simulation-based evaluations are summarized in Sections VIII-A3 and VIII-A4.

1) Real World Datasets: Table XIX provides a summary of publicly available datasets for CP. These datasets predominantly focus on vehicle-to-infrastructure (V2I) collaboration, with limited attention given to vehicle-to-vehicle (V2V) interactions. The absence of datasets that integrate both V2I and V2V collaborations highlights a significant gap in existing resources, underscoring the need for more comprehensive datasets to advance CP research.

LiDAR emerges as the most frequently used sensor in these datasets, often complemented by RGB cameras to enhance visual perception. However, the exclusion of additional modalities, such as infrared cameras or radars, limits their utility in handling complex scenarios, particularly in adverse environmental conditions. Furthermore, the scale of these datasets remains small compared to single-entity perception datasets like NuScenes [122] and Waymo [123], which feature over 200,000 frames. This disparity is further compounded by limited scenario diversity, as most datasets are constrained to daytime and clear weather conditions. To address these shortcomings, future datasets should incorporate a wider variety of scenarios, including nighttime and adverse weather, to better reflect real-world challenges in CP.

The diversity in annotated object classes across these datasets also reveals notable inconsistencies. For example, DAIR-V2X [124] includes annotations for 10 distinct object classes, whereas others, such as [125] and [126], focus on fundamental categories like pedestrians, cyclists, cars, and trucks. Some datasets adopt a task-specific approach, such as [88], which is exclusively dedicated to pedestrian detection. Although object detection is a central feature, support for advanced tasks like object tracking is limited, and motion prediction remains under-represented, highlighting an imbalance in task coverage.

Table XX delves deeper into V2X configurations, examining critical aspects such as the number of connected vehicles, localization methods, and time synchronization protocols. Most datasets involve three or fewer CAVs, as exemplified by LUCOOP [126]. Time synchronization is generally asynchronous, with a maximum latency of 50 milliseconds between entities. To ensure accurate ground-truth localization, hybrid localization approaches are commonly employed, combining multiple techniques such as HD Map and Real Time Kinematic (RTK) to minimize positional errors. These precise localization methods play a pivotal role in enhancing CP system performance and fostering effective collaboration between vehicles and infrastructure.

2) Synthetic Datasets: Table XXI summarizes the synthetic datasets used in CP. These datasets are predominantly generated using CARLA, often paired with frameworks like OpenCDA [129], which integrates CARLA with SUMO for

traffic simulation. The use of simulation significantly reduces the effort required to create scenarios involving multiple CAVs compared to real-world settings. Unlike real-world datasets, synthetic datasets frequently support both V2V and V2I communication, making them highly versatile for CP research. Additionally, many synthetic datasets enable cooperation among more than three CAVs, further enhancing their applicability. These datasets also exhibit greater diversity in sensor modalities, incorporating LiDAR and RGB cameras, with some extending to include depth information.

Synthetic datasets exhibit greater diversity in the range of tasks they support. While object detection remains the primary focus, many datasets extend their scope to include tasks such as semantic segmentation and accident prediction. However, the scenario diversity is often constrained by a reliance on pre-existing maps from CARLA, which limits geographic variety and reduces the ability to replicate a wide range of real-world conditions accurately.

A significant characteristic of synthetic datasets is their reliance on idealized system conditions. As outlined in Table XXII, most datasets assume perfect time synchronization between connected entities, with the exception of [48]. Additionally, simulators provide precise ground-truth localization, resulting in error-free localization performance. While these conditions simplify evaluation, they may not fully reflect the challenges of real-world scenarios.

Despite these limitations, synthetic datasets are often designed to replicate real-world conditions to better evaluate CP system performance. For instance, [103] explores the impact of time latency and pose errors, demonstrating that increasing the number of CAVs enhances robustness against such issues. Similarly, [130] reveals a positive correlation between the number ofCAVs and the average precision of object detection, with performance improvements plateauing at around four CAVs. These studies highlight the value of synthetic datasets in examining trade-offs and identifying limitations in CP systems.

3) Real World Experiment: As discussed in the previous section, existing real-world datasets have significant limitations. When new algorithms offer advantages that cannot be effectively demonstrated using these datasets, dedicated experiments become essential. However, real-world experiments demand considerable time and financial resources, making them far less common than simulation-based evaluations.

For example, Sakr et al. [107] conduct an experiment where a legacy vehicle follows a sensor-rich vehicle that transmits road geometry information. The aim is to estimate the road geometry ahead of the legacy vehicle using the data provided by the sensor-rich vehicle. Similarly, Li et al. [66] design experiments involving two vehicles equipped with LiDARs and cameras. Their study demonstrates two scenarios where V2V perception outperforms single-entity perception, particularly in detecting distant objects beyond the range of LiDAR sensors. Additionally, they show how V2V communication effectively reduces positioning errors in various road scenarios. Xie et al. [35] adopt a different approach by conducting real-world experiments with two vehicles equipped with LiDARs and cameras. These vehicles collect data across three representative V2V scenarios, facilitating the validation of their algorithm under real-world conditions.

In general, real-world experiments remain rare due to the significant resources required. Most of these studies focus on V2V cooperation, as it is easier to design and continues to be the most extensively researched form of V2X collaboration.

4) Simulation Experiment: In simulation experiments, trends similar to those observed in real-world evaluations are evident. Table XXIII summarizes studies employing simulation-based experiments, which either generate new datasets or adapt existing ones to evaluate specific approaches, as demonstrated in [14].

As with real-world tests, V2V communication remains the dominant approach, preferred over V2I or combined V2I and V2V methods. Object detection continues to be the most frequently studied perception task [8], [10], [11], [37], [38], [58], [65], [76], [81], [132]. Some studies create tailored datasets to address specific requirements, such as semantic segmentation [92], [95] or lane detection [108].

CARLA[4] is the most widely used simulator for CP research, frequently utilized in customized configurations. AirSim[5] and Gazebo[6] are also commonly employed. Among these, only one study incorporated a network simulator to model realistic network data traffic and its impact on perception performance [10].

Simulation studies often explore specific aspects of CP. A recurring focus is determining the optimal number of cooperating vehicles to maximize performance [10], [14], [102]. Kuang et al. [81] investigate scenarios where V2V cooperation significantly outperforms single-vehicle perception. Similarly, Liu et al. [58] analyze the effects of homogeneous versus heterogeneous sensor configurations on CP performance across various conditions.

Another line of research examines federated learning for CP. For instance, Zhang et al. [13] implement a dynamic map fusion algorithm using federated learning to recover objects missed by individual systems, demonstrating its potential to enhance CP performance.

B. Evaluation Scenarios

1) Environmental Settings: The methodologies for evaluating CP algorithms, including datasets and experiments, were introduced in the previous section. This section examines the specific scenarios used for algorithm evaluation. In both real-world and simulation studies, environments are typically categorized into three primary types: urban, rural, and highway, as outlined in Table XXIV. Simulation studies offer a wider range of scenarios compared to real-world evaluations, largely due to the extensive use of CARLA and its pre-defined maps. However, this reliance on CARLA maps introduces limitations in the diversity of road environments and the assessment of specific road features, such as intersections. Although many studies specify which CARLA map is utilized, detailed information about the types of intersections or the

[4]https://carla.org
[5]https://microsoft.github.io/AirSim
[6]https://gazebosim.org

specific routes examined is often absent. Commonly evaluated road configurations, including cross intersections and straight road segments, are summarized in Table XXV.

2) Daytime and Weather: Robust evaluation of CP algorithms requires testing under diverse conditions to assess their performance in challenging scenarios, such as low-light environments or adverse weather conditions. Many real-world datasets incorporate both daytime and nighttime data [99], [126], [128], although not all studies provide explicit documentation of these conditions [125], [127]. Regarding weather diversity, detailed descriptions are frequently omitted. Among real-world datasets, DAIR-V2X [124] stands out for its inclusion of varying weather and lighting conditions, establishing it as the most comprehensive dataset in this regard.

In contrast, synthetic datasets offer complete control over environmental parameters such as weather and time of day. However, these conditions are rarely detailed in the associated studies. An exception is the DeepAccident dataset [103], which explicitly provides variations in weather conditions (e.g., clear, rainy, cloudy, wet) and times of day (e.g., noon, sunset, night). This level of specification enhances its utility for evaluating CP algorithms under diverse environmental settings.

C. Evaluation Metrics

To quantitatively assess perception performance, various metrics are applied to specific tasks such as object detection, tracking, and motion prediction, depending on the evaluation objectives. Unified evaluation metrics are crucial for benchmarking different algorithms, enabling comparative analysis of their performance, and supporting the continuous improvement of these algorithms.

This section reviews and summarizes the metrics used for evaluating CP. These metrics are categorized into two groups: general evaluation metrics, which are adapted from single-entity perception tasks, and custom metrics designed for CP. Additionally, this section provides a summary of the ablation studies conducted in the reviewed papers. These studies offer insights into the evaluation process, highlighting common factors that impact CP and how they influence performance. This understanding aids researchers in designing more practical and robust CP frameworks.

1) General Evaluation Metrics for Perception Tasks: The general evaluation metrics for different perception tasks are summarized in the Table XXVI. These metrics are adapted from single-entity perception and are widely accepted by researchers. In some cases, evaluation results are divided into different groups based on detection difficulty levels, such as easy, medium, and difficult, as seen in the KITTI dataset [133], which considers factors like occlusion level and object size. Evaluation results can also be categorized by object type, such as cars, cyclists, and pedestrians. This categorization helps researchers better understand the strengths and limitations of different approaches.

2) Custom Evaluation Metrics for Collaborative Perception: Traditional evaluation metrics used for single-entity perception do not adequately represent the performance of cooperative perception (CP). Since CP primarily aims to address visual occlusion problems and serves as a supplement to single-entity perception, it requires distinct evaluation criteria. Moreover, CP is significantly constrained by communication resources. Therefore, communication factors should be incorporated into the design of evaluation metrics. In this subsection, we summarize custom metrics designed for CP in Table XXVII. These metrics are classified into the following three types.

- **Communication**: Compared to single-entity perception, cooperative perception requires additional communication resources. Evaluating the communication demands of CP algorithms is crucial for assessing their efficiency and scalability. For example, metrics such as average message size are commonly used to measure the communication costs associated with CP.
- **Perception**: CP aims to address visual occlusion problems, making it crucial to have metrics that assess how effectively CP resolves these issues. Wang et al. [114] introduce the Average Recall of Collaborative View (ARCV) metric, which measures the average recall of agents that are invisible from a single-vehicle perspective but become detectable through collaborative perception. In addition to uncovering occluded agents, CP can enhance the perception of agents already visible to a single vehicle by incorporating additional information. To quantify this enhancement, Luo et al. [43] propose the marginal gain metric, defined as the performance improvement when an additional agent joins the collaboration. It is important to note that the marginal gain tends to diminish as more observing agents are added.
- **Ratio between Communication and Perception**: There is an inherent trade-off between communication cost and collaborative perception (CP) performance. Reducing communication costs can constrain CP effectiveness by limiting the amount of shared information among agents. Researchers are exploring how to balance these factors to develop efficient and effective CP approaches. For instance, Liu et al. [91] introduce the Bandwidth Improvement Score (BIS), defined as the ratio of the relative improvement in overall accuracy to the bandwidth usage. A higher BIS indicates a more favorable balance, lower bandwidth cost coupled with greater improvement in perception performance.

The custom metrics for collaborative perception place greater emphasis on improving the detection of both visible and previously invisible objects from the ego vehicle's viewpoint. However, these perception improvement metrics have not been widely accepted in the research community. Most studies predominantly utilize evaluation metrics adopted from single-entity perception, with the exception of studies [33], [50], [91], [114], which employ custom metrics for CP. Communication cost metrics are also occasionally considered when evaluating the efficiency of collaborative perception methods.

3) Ablation Studies: Ablation studies are crucial for evaluating the robustness and scalability of CP systems under various conditions. They help identify how different

factors affect CP performance, enabling researchers to optimize system design. In this section, we categorize and discuss ablation studies focusing on communication, localization, system scale, visual occlusion, adversarial attacks, and other relevant factors.

- **Communication**: Communication poses significant challenges for CP in real-world applications. Vehicle-to-everything (V2X) communication introduces practical constraints such as bandwidth limitations, data compression requirements, latency, noise, and interruptions. Most research includes ablation studies addressing these communication issues [8], [9], [15], [36], [50], [54], [69], [75], [91]. It has been proven that latency and noise can significantly degrade CP performance, while interruptions have the most severe impact. Designing communication-aware approaches is essential for enhancing the scalability and effectiveness of CP applications.
- **Localization**: Localization errors heavily affect the performance of CP systems. To ensure algorithms are viable in real-world settings, it is crucial to measure their robustness against positioning errors. Most studies validate algorithm performance under varying positioning errors, typically ranging from 0 to 1 meter [15], [39], [46], [51], [64], [71], [72], [74], [77], [91]. The absence of significant performance degradation under these conditions demonstrates the algorithm's reliability in practical applications.
- **Scale of system**: The performance of CP varies with the number and types of CAVs involved. Ablation studies are helpful in determining the optimal configuration of cooperative systems. Research has shown that CP achieves the best results when 4 to 6 agents participate in the collaborative system; adding more agents does not further increase perception accuracy [34], [44], [60], [64]. Additionally, the types of CAVs, such as those equipped with LiDAR or cameras, also influence CP performance [22]. Validating CP under different ratios of LiDAR-equipped and camera-equipped CAVs is important to ensure robustness.
- **Visual occlusion**: Verifying CP's reliability in detecting occluded objects requires ablation studies that consider different levels of occlusion. These studies demonstrate the effectiveness of CP in addressing visual occlusion and indicate how well the system performs under such conditions [13], [100].
- **Adversarial attack**: Robustness against adversarial attacks is a critical aspect of CP systems. Ablation studies focusing on attack scenarios verify whether CP can maintain reliability under various adversarial conditions [23], [61]. Ensuring resilience to such attacks is vital for the safe deployment of CP systems.
- **Others**: Additional factors can affect CP performance, such as traffic density, vehicle velocity and speed, and sensor dropout. Conducting diverse ablation studies under different scenarios ensures the system's reliability in real-world usage [9], [94]. By validating CP performance across these variables, researchers can develop more robust and adaptable systems.

Various ablation studies have been conducted to assess the reliability and robustness of CP systems. However, performing comprehensive ablation studies is time-consuming and resource-intensive. Researchers should prioritize validating factors that are most pertinent to the specific problems their work aims to address. To simplify the evaluation process, an automated evaluation framework is needed.

While ablation studies are valuable for measuring CP performance, they may not always yield accurate results due to the interplay of multiple influencing factors, such as the number of CAVs and communication bandwidth. To ensure reliable validation, it is important to conduct online evaluations using simulations or real-world experiments. These methods can address the interdependencies of various factors, filling the gap left by traditional ablation studies.

IX. Challenges, Opportunities, and Risks (RQ5)

Collaborative Perception holds significant potential to extend the perception range of individual vehicles and address critical scenarios caused by occlusion. However, implementing this technology in real-world applications faces numerous challenges. Based on the comprehensive analysis of CP, this section introduces the challenges, opportunities, and risks associated with CP research.

We examine the challenges and opportunities from three perspectives: hardware, software, and evaluation methods. The risks in CP research are summarized concerning application gaps, reproducibility, and evaluation. Each aspect provides insight into the current state of CP and highlights areas for future improvement.

A. Challenges

1) Hardware: CAVs employ a variety of sensors, each with its advantages and limitations. These vehicles are typically equipped with multiple sensors, such as LiDAR and cameras, to navigate diverse driving scenarios effectively. Integrating these sensors enhances the capabilities of multi-modality in CP, allowing vehicles to perceive their environment more comprehensively and share multi-modal information with nearby vehicles. However, achieving precise time synchronization and calibration among multiple sensors is challenging. Multi-modal CP methods rely on accurate spatial and temporal alignment from different sensors, but factors like sensor drift and environmental variability make consistent precision challenging to maintain. This inconsistency can hinder the full potential of multi-modal approaches. To fully harness the advantages of multi-modality, it is essential to develop efficient calibration methods for multi-sensor systems, not only on the vehicles themselves but also within the supporting infrastructure [25]. Addressing these calibration challenges will enhance the reliability and effectiveness of CP in real-world applications.

2) Software: While hardware challenges such as sensor calibration are significant, various software challenges also exist and are the main focus of this literature review. In the following sections, we discuss these software challenges from two critical aspects, communication and perception, which together form the core technologies of cooperative perception.

- **Communication**: V2X communication enables data transmission between entities but comes with certain constraints. As discussed in Section VII, the communication challenges involved in CP are significant. Bandwidth limitations and communication range constraints are primary considerations when designing a CP framework. To prevent network congestion, the framework must minimize bandwidth demands. In addition to communication efficiency, the robustness of the CP framework against latency, data loss, and interruptions is crucial for maintaining reliable perception. By addressing these real-world communication factors, we can effectively implement CP technology in practical applications, enhancing perception in critical scenarios and improving the safety of CAVs. However, only one study, V2X-INCOP [20], has specifically addressed communication interruptions. Research on robust CP under realistic communication conditions remains significantly under-explored. In addition to addressing communication constraints, the standardization of V2X protocols presents significant challenges for early and intermediate collaboration approaches. Currently, standardized CP exclusively support late collaboration. Exploring effective methods to transmit raw sensor data and intermediate features within the framework of realistic communication protocols remains a critical area of investigation.
- **Fusion strategy in Perception**: Information fusion among agents is central to CP, enabling a collective understanding of the environment. However, several challenges persist in developing efficient and robust fusion methods. Firstly, information loss is a significant concern in data fusion. Techniques such as late fusion, which combine perception results using bounding boxes, often discard crucial texture information. Traditional feature fusion methods, such as average pooling, may overlook detailed features from different agents. These fine-grained details are essential for accurate scene understanding. To overcome these limitations, exploring efficient data fusion methods that retain essential information for downstream tasks without substantially increasing communication costs is necessary. Secondly, the growing volume and variety of data shared among agents introduce challenges in data management and resource allocation. Novel hybrid fusion methods that utilize features and perception results can enhance cooperation between agents, such as Hybrid-CP [64]. However, the inclusion of diverse data types substantially increases the complexity of data management. Managing heterogeneous data from multiple agents poses a significant challenge, necessitating targeted solutions. Lastly, data alignment remains a bottleneck in the real-world application of CP systems. Spatial alignment issues are not fully resolved; most current approaches are only robust against positioning errors within one meter [19], [20], [48], [52], [99], [114], [115]. In practice, the localization error of CAVs can vary by several meters. Achieving higher robustness is essential to scale CP solutions across diverse conditions. For instance, resolving the positional alignment of visual features extracted from different agents' cameras remains an unsolved problem. Temporal alignment is another crucial factor. Aligning asynchronous features is particularly challenging because it often relies on predicting future features, which can constrain the accuracy of the alignment.
- **Robustness**: Ensuring the robustness of CP systems is crucial for autonomous driving applications, which are inherently safety-critical. Various factors can degrade the performance of CP systems. As previously discussed, issues with communication and localization significantly affect performance, making it essential to enhance the robustness of CP systems against these challenges. In addition to these factors, challenging scenarios present further critical issues. For example, collaborative lane detection performance in complex road structures deteriorates because current models may not be sufficiently robust to infer intricate road geometries accurately in real-time [107], [108], [109]. Similarly, collaborative object detection performance declines in dense traffic conditions due to increased occlusion and the constrained bandwidth available to each CAV in the area. Although CP has not yet been extensively evaluated under adverse weather conditions, performance is expected to degrade similar to single-entity perception systems. Therefore, enhancing the robustness of CP systems across diverse scenarios is imperative. Beyond environmental challenges, adversarial attacks pose another significant threat to CP. With increasing connectivity between vehicles, infrastructure, and cloud services, protecting autonomous vehicles from network attacks becomes more critical. CP systems should be capable of identifying fraudulent messages and avoiding the fusion of malicious data to maintain system reliability. Only one study, AmongUs [23], implement the detection of malicious activities within the CP framework and evaluated their impact on perception performance. This remains a significantly under-explored area.
- **Uncertainty**: CP relies heavily on Artificial Intelligence (AI), which often functions as a "black box" due to its lack of explainability. This opacity makes it difficult to determine absolute confidence in the perception results. CAVs from different manufacturers may use diverse perception models with varying performance levels. In CP systems, receivers obtain processed data from senders – such as detected objects – but assessing the uncertainty associated with this data is challenging. This situation raises the issue of trustworthiness: to what extent should CAVs trust the information received from others? Beyond uncertainties in perception results, there are inherent uncertainties within the systems. For example, depth estimation using cameras for 3D perception introduces uncertainty, as do upstream tasks whose errors can accumulate throughout the processing pipeline, ultimately degrading the outcome. To enhance the reliability and explainability of CP systems, it is important to design uncertainty-aware models that can learn to process noisy data effectively.
- **Efficiency**: Autonomous driving systems are computationally intensive platforms that process large amounts

of data in real-time. Perception is one of the most resource-consuming modules, heavily relying on complex neural networks. Compared to single-entity perception, CP demands even more computing and communication resources. The trade-off between improved perception and additional resource consumption is a critical factor in determining the scalability of CP systems for real-world applications. Furthermore, real-time performance necessitates that CP systems achieve computational and communication efficiency. This ensures that accurate information is transmitted promptly to downstream modules such as planning and control. Balancing these demands is essential for effectively deploying CP in practical autonomous driving scenarios.

- **Domain shift**: Domain shift presents a significant challenge in CP among agents equipped with different types of LiDAR sensors. Features extracted from various LiDAR systems do not reside in the same feature space, meaning this discrepancy can significantly degrade system's performance [21], [58]. To address this issue, bridging the gap between the source and target domains is essential before fusing features from multiple agents. Beyond its impact on feature fusion, the simulation-to-real-world (Sim2Real) domain shift also causes models trained on synthetic datasets to perform poorly in real-world environments [57], [62]. Collecting real-world data is costly and particularly difficult for safety-critical scenarios. As a result, researchers and developers are seeking cost-efficient solutions for training neural networks using synthetic data. However, the pronounced gap between simulation and reality makes achieving this challenging. To increase the utilization of synthetic data in CP, it is urgent to bridge the Sim2Real gap, facilitating the transfer of knowledge learned from simulations to real-world applications. This advancement would also enable training models with synthetic safety-critical data, filling current gaps in available training datasets.
- **Heterogeneity**: Research on CP often assumes unrealistic conditions to simplify the complexity of collaborative systems, particularly regarding the heterogeneity of agents. This heterogeneity includes differences in models (model heterogeneity) [121] and sensing modalities (modality heterogeneity) [22], [77], [80]. Embracing heterogeneous collaboration is essential for making CP technology applicable in industry and deployable in real-world scenarios. However, current approaches that address heterogeneity are limited and struggle to maintain the reliability of CP systems under heterogeneous conditions.
- **Model training**: Model training is a crucial step in developing CP algorithms. As models increase in size and complexity, they require larger datasets for effective training. To reduce costs, it is important to decrease the dependence on labeled data in CP, which can significantly reduce the effort required for data annotation [26].

3) Evaluation: The evaluation of CP systems presents several challenges, ranging from the methods used to the metrics applied. The challenges related to evaluation, as identified in the literature, are summarized below.

- **Lack of large-scale real-world dataset**: AI-driven perception algorithms require large-scale and diverse datasets to learn the patterns and features necessary for robust model generalization. However, the current datasets available for CP research are insufficient in size and lack diversity. They do not adequately cover a range of scenarios, such as different weather conditions or critical traffic situations. Additionally, existing datasets primarily support collaborative object detection, tracking, and prediction, but there are no datasets for tasks like collaborative semantic segmentation or lane detection. To advance CP research forward, creating large-scale, multi-modal datasets that support multiple tasks across diverse scenarios is essential. Creating a real-world dataset presents several challenges that must be addressed in advance. Data privacy concerns and the processes required to ensure compliance can be time-intensive. In particular, visual data captured by cameras must undergo anonymization to obscure identifiable features such as human faces and vehicle license plates, ensuring adherence to data privacy regulations. Hardware setup poses additional difficulties, particularly in achieving precise time synchronization and localization for the vehicles involved. Moreover, generating diverse annotations, covering supported tasks, object classes, or supplementary details such as occlusion, demands substantial time and financial resources. Finally, as these datasets are typically recorded at real-world intersections rather than controlled test fields, managing class distribution becomes challenging due to the lack of control over traffic conditions.
- **Simulation for evaluation**: To further advance CP algorithms, designing fair and goal-oriented evaluation methods that quantitatively measure their performance is essential. Beyond benchmarking on public datasets, conducting online evaluations in simulations that consider more realistic network conditions can provide deeper insights. However, integrating realistic communication models into co-simulation frameworks remains challenging due to bottlenecks between multiple simulation platforms.
- **Scenarios for evaluation**: CP is designed to address critical occlusion situations to enhance the safety of CAVs. However, collecting data on these critical scenarios from real-world environments is challenging. Such situations are rare in daily traffic and pose significant risks during data collection. Consequently, there is a gap in validating the reliability of CP in safety-critical scenarios, which is crucial for its improvement. Developing effective methods to evaluate CP under these conditions remains an essential yet unresolved challenge.
- **Evaluation metrics**: Metrics are essential for quantitatively analyzing the performance of CP methods. Therefore, it is crucial to design appropriate metrics that clearly represent the advantages and limitations of CP. Most existing metrics are adapted from single-entity perception, which may not fully capture the unique benefits of CP, especially in addressing occlusion. Notably, only

one study, UMC [114], evaluates performance specifically on occluded objects. To effectively evaluate CP's efficiency in solving occlusion problems, new metrics need to be developed, introducing novel criteria for quantitative assessment. Beyond performance measurement, there is also a need for metrics that evaluate effectiveness and safety-related aspects. Developing such metrics will enable a more comprehensive and quantitative analysis of CP systems, facilitating their improvement and real-world application.

- **Ablation study**: Researchers employ ablation studies to evaluate the efficiency of CP under various conditions. However, conducting these studies is more labor-intensive than in single-entity perception due to the diverse factors affecting CP, including communication, localization, and perception. To accelerate future research, it is essential to develop tools that enable the automatic execution of ablation studies.

B. Opportunities

We have outlined corresponding opportunities and future directions by identifying the open challenges and research gaps in CP. These are summarized from three critical perspectives: hardware, software, and evaluation.

1) Hardware:

- **Optimal sensor configuration**: Optimizing sensor configurations is a significant opportunity in the hardware aspect of CP. Researchers have ample potential to explore which hardware setups are most effective for CP systems. Determining the optimal types and placements for infrastructure sensors is particularly important. The design and positioning of sensors at intersections directly impact roadside perception performance [134]. Investigating the trade-offs between sensor redundancy and safety is also valuable for enhancing system reliability.
- **New modality**: Another area for advancement is the integration of new sensor modalities. Current CP frameworks predominantly use LiDAR and cameras to perceive the environment. However, the application of radar, infrared cameras, or event cameras remains largely unexplored. Radars can provide more accurate velocity measurements, while infrared cameras offer night vision capabilities. Incorporating these sensors can enhance the robustness of perception systems by supplementing the limitations of LiDAR and standard cameras.

2) Software:

- **Communication**: Enhancing communication efficiency is a critical opportunity in the software aspect of CP. Since communication is fundamental to these systems, improving it can significantly boost overall performance. One approach is implementing data compression techniques that reduce message sizes without substantial information loss. This applies to raw data and feature data, enabling the transmission of more valuable information and enhancing the CP process. Additionally, exploring the transmission of various data types, maps and historical perception data, can diversify CP solutions. Designing efficient data structures for these heterogeneous data types is crucial for real-world applications. Implementing contributor selection strategies can also reduce unnecessary connections and data redundancy within the collaborative framework [17], [50], [91].
- **Fusion strategy in CP**: Advancing fusion strategies presents another promising direction. Hybrid fusion methods have shown potential in balancing resource consumption with perception performance by dynamically adapting to communication conditions, thus ensuring scalability [35], [61], [64], [65], [87]. Further research into hybrid fusion could unlock more benefits for CP. Graph-based feature fusion is an underexplored area that merits attention. Graph Neural Networks (GNNs) can model agent relationships and adjust collaborations based on changing environments. Investigating the application of GNNs in communication and perception could yield significant advancements.
- **Robustness and efficiency**: Improving robustness and efficiency is vital for the practical deployment of CP systems. While issues like communication disruptions and adversarial attacks have been studied, hardware failures, such as sensor dropouts, have been largely overlooked. Enhancing robustness against sensor failures is important for ensuring system reliability. On the efficiency front, although many state-of-the-art algorithms achieve high accuracy on public datasets, their performance concerning hardware limitations has not been thoroughly investigated. Exploring ways to improve algorithmic efficiency, such as optimizing models like V2X-ViT [16], would be a valuable direction for future research.
- **Compatibility**: Lastly, ensuring compatibility between CP and ego perception systems is essential. CP should supplement, not replace, individual perception capabilities. Current research often treats CP as a separate system, leading to potential resource wastage. Developing perception pipelines that can operate independently without shared information and collaborate with other agents when necessary would make systems more practical. Designing CP as a plug-and-play module can avoid the need for complete redesign and retraining of perception models, enhancing efficiency and adaptability.

3) Evaluation: Evaluation methods are essential guidelines for researchers and developers seeking to improve CP performance. However, current evaluation approaches have notable drawbacks. This section outlines opportunities to enhance CP evaluation from several perspectives.

- **Datasets**: A significant opportunity lies in developing large-scale CP datasets to advance research. Besides size, diversity in datasets is equally important. Incorporating various sensor modalities and perception tasks can enrich future CP datasets. While annotating real-world data is costly, researchers might provide unlabeled data to the community to foster collaboration. Creating an open-source framework for generating synthetic CP datasets would also be highly beneficial, enabling broader participation and innovation.

- **Evaluation methods**: To bridge the deployment gap in CP systems, developing a framework simplifying the entire lifecycle, from research and development to deployment and testing, is crucial. Such a framework should accelerate validation and evaluation with datasets and in real-world conditions. By streamlining these processes, CP can more readily transition into practical applications.
- **Evaluation scenarios**: Validating the reliability of CP under diverse conditions requires collecting a more comprehensive range of evaluation scenarios. Expanding the diversity of these scenarios ensures that CP systems are robust across different environments. Particular attention should be given to critical traffic situations, such as those involving vulnerable road users, to assess system's performance in high-risk contexts thoroughly.
- **Metrics and ablation studies**: Quantitative metrics are important for accurately measuring CP performance. As discussed in the challenges, new metrics that align with CP's objectives, such as resolving visual occlusions, are needed. Beyond developing new metrics, creating a framework that enables automatic ablation studies under varied conditions would provide valuable insights. Such a framework can help researchers understand the impact of different components and configurations, ultimately leading to more effective CP systems.

C. Risks

While CP technology shows great promise in enhancing the capabilities of CAVs and improving road safety, it also faces several significant risks.

- **Deployment gap**: A significant risk is the gap between research advancements and real-world deployment. Although various studies address individual aspects of this gap, the complexities of real-world conditions far exceed those modeled in datasets or simulations. For instance, limited communication bandwidth caused by environmental factors, unexpected synchronization failures, as well as unpredictable communication latency and interruptions can adversely affect CP performance. Additionally, striking the right balance between the perceptual improvements gained through collaboration and the additional costs incurred is challenging. Successful deployment of CP also requires collaboration among different vendors to ensure that CAVs from various manufacturers can communicate effectively and have compatible perception modules.
- **Reproducibility**: Another critical concern is the reproducibility of research findings. The research community must verify results and build upon previous work to ensure reproducibility. However, the limited availability of accessible repositories and source code in CP research hampers this process. Providing open-source code and datasets is highly encouraged to enable other researchers to reproduce results and advance the field.

Despite these risks, the substantial potential benefits of CP make it a valuable area for continued exploration. Overcoming these challenges will require collaboration among researchers from various disciplines, including computer vision, communication technology, and vehicle engineering. Companies and governments should actively work together to establish standards for different V2X applications and develop compatible CP systems. Finally, embracing open-source practices can significantly assist the research community in reproducing results and focusing on new challenges.

X. Conclusion

In this paper, we systematically reviewed recent research on Collaborative Perception (CP). We propose a structured taxonomy categorized by modality, collaboration type, and task, encompassing object detection, tracking, motion prediction, segmentation, lane detection, and multi-task or task-agnostic pipelines. The review also examined advanced techniques addressing real-world challenges, including pose errors, latency, bandwidth limitations, communication interruptions, domain shifts, heterogeneity, and adversarial attacks. Furthermore, we conducted a comparative analysis of these approaches, highlighting their strengths and limitations, and reviewed CP evaluation methods, ranging from real-world datasets and synthetic datasets to experiments in real-world and simulated environments. Key limitations in current evaluation scenarios and metrics were identified, alongside challenges and opportunities in hardware, software, and evaluation methodologies.

A central motivation for CP is to address visual occlusions and complement ego-perception systems. However, current research often overlooks the necessity of ensuring compatibility between CP and ego-perception pipelines, as well as the importance of triggers to selectively activate collaboration under appropriate conditions. To assess CP's effectiveness in addressing visual occlusions, novel evaluation approaches aligned with its goals are essential. This review underscores the urgent need for large-scale CP datasets that reflect realistic setups and diverse scenarios, which are pivotal for advancing the field.

Future work must prioritize the development of appropriate evaluation methodologies and large-scale datasets. An open-source co-simulation framework that represents realistic real-world scenarios and a unified collaborative driving framework encompassing the entire lifecycle, from research and development to deployment and validation, could significantly accelerate CP advancements and its real-world implementation. Bridging the deployment gap should remain a key focus for future investigations.

Through this systematic review, we re-evaluate the concrete role of CP in CAVs. Revolutionizing evaluation methods and addressing deployment challenges will help transition CP systems from lab prototypes to real-world applications. As CP systems integrate communication, vehicular, and computer vision technologies, their progress will require interdisciplinary collaboration to enable the practical deployment of sophisticated CP solutions.

Given the time constraints of this survey, our literature collection was finalized in March 2024, while our review focuses on research published in the past five years (2019–2023). However, our systematic review protocol and curated paper

set provide a solid foundation for future researchers to extend this study. By applying forward snowballing, researchers can efficiently update the review with high-quality, cutting-edge research beyond our collection period.

ACKNOWLEDGMENT

The authors are solely responsible for the content of this publication. Figure 1 was created with the Car2Car Communication Consortium (C2C-CC) illustration toolkit.

REFERENCES

[1] 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 *Proc. 4th ACM/IEEE Symp. Edge Comput.*, Nov. 2019, pp. 88–100, doi: 10.1145/3318216.3363300.

[2] H.-J. Günther, O. Trauer, and L. Wolf, "The potential of collective perception in vehicular ad-hoc networks," in *Proc. 14th Int. Conf. ITS Telecommun. (ITST)*, Dec. 2015, pp. 1–5, doi: 10.1109/ITST.2015.7377190.

[3] G. Thandavarayan, M. Sepulcre, and J. Gozalvez, "Generation of cooperative perception messages for connected and automated vehicles," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16336–16341, Dec. 2020, doi: 10.1109/TVT.2020.3036165.

[4] Q. Delooz et al., "Analysis and evaluation of information redundancy mitigation for V2X collective perception," *IEEE Access*, vol. 10, pp. 47076–47093, 2022, doi: 10.1109/ACCESS.2022.3170029.

[5] Q. Delooz, R. Riebl, A. Festag, and A. Vinel, "Design and performance of congestion-aware collective perception," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Dec. 2020, pp. 1–8, doi: 10.1109/VNC51378.2020.9318335.

[6] *Intelligent Transport Systems (ITS); Collective Perception Service (CPS); Release 2*, document TS 103 324, ETSI, 2023.

[7] SAE, *V2X Sensor-Sharing for Cooperative and Automated Driving*, document J3224_202208, Aug. 2022.

[8] A. N. Ahmed, I. Ravijts, J. de Hoog, A. Anwar, S. Mercelis, and P. Hellinckx, "A joint perception scheme for connected vehicles," in *Proc. IEEE Sensors*, Oct. 2022, pp. 1–4, doi: 10.1109/SENSORS52175.2022.9967271.

[9] Q. Zhang, X. Zhang, R. Zhu, F. Bai, M. Naserian, and Z. M. Mao, "Robust real-time multi-vehicle collaboration on asynchronous sensors," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2023, pp. 1–15, doi: 10.1145/3570361.3613271.

[10] G. Luo et al., "EdgeCooper: Network-aware cooperative LiDAR perception for enhanced vehicular awareness," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 1, pp. 207–222, Jan. 2024, doi: 10.1109/JSAC.2023.3322764.

[11] E. E. Marvasti, A. Raftari, A. E. Marvasti, and Y. P. Fallah, "Bandwidth-adaptive feature sharing for cooperative LiDAR object detection," in *Proc. IEEE 3rd Connected Automated Vehicles Symp. (CAVS)*, Nov. 2020, pp. 1–7, doi: 10.1109/CAVS51000.2020.9334618.

[12] E. E. Marvasti, A. Raftari, A. E. Marvasti, Y. P. Fallah, R. Guo, and H. Lu, "Cooperative LiDAR object detection via feature sharing in deep networks," in *Proc. IEEE 92nd Veh. Technol. Conf. (VTC-Fall)*, Nov. 2020, pp. 1–7, doi: 10.1109/VTC2020-Fall49728.2020.9348723.

[13] Z. Zhang, S. Wang, Y. Hong, L. Zhou, and Q. Hao, "Distributed dynamic map fusion via federated learning for intelligent networked vehicles," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 953–959, doi: 10.1109/ICRA48506.2021.9561612.

[14] G. Volk, J. Gamerdinger, A. von Bernuth, S. Teufel, and O. Bringmann, "Environment-aware optimization of track-to-track fusion for collective perception," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2385–2392, doi: 10.1109/ITSC55140.2022.9922388.

[15] Z. Song, F. Wen, H. Zhang, and J. Li, "A cooperative perception system robust to localization errors," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–6, doi: 10.1109/IV55152.2023.10186727.

[16] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 107–124, doi: 10.1007/978-3-031-19842-7_7.

[17] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4105–4114, doi: 10.1109/CVPR42600.2020.00416.

[18] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun, "Learning to communicate and correct pose errors," in *Proc. Conf. Robot Learn. (CoRL)*, Oct. 2021, pp. 1195–1210. [Online]. Available: https://proceedings.mlr.press/v155/vadivelu21a/vadivelu21a.pdf

[19] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Proc. 17th Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 316–332, doi: 10.1007/978-3-031-19824-3_19.

[20] S. Ren et al., "Interruption-aware cooperative perception for V2X communication-aided autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 4, pp. 1–17, Apr. 2024, doi: 10.1109/TIV.2024.3371974.

[21] X. Li, J. Yin, W. Li, C. Xu, R. Yang, and J. Shen, "DI-V2X: Learning domain-invariant representation for vehicle-infrastructure collaborative 3D object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, Dec. 2024, pp. 3208–3215, doi: 10.1609/aaai.v38i4.28105.

[22] H. Xiang, R. Xu, and J. Ma, "HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 284–295, doi: 10.1109/ICCV51070.2023.00033.

[23] Y. Li, Q. Fang, J. Bai, S. Chen, F. Juefei-Xu, and C. Feng, "Among us: Adversarially robust collaborative perception by consensus," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 186–195, doi: 10.1109/ICCV51070.2023.00024.

[24] Z. Bai et al., "A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 15191–15209, Nov. 2024, doi: 10.1109/TITS.2024.3436012.

[25] A. Caillot, S. Ouerghi, P. Vasseur, R. Boutteau, and Y. Dupuis, "Survey on cooperative perception in an automotive context," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14204–14223, Sep. 2022, doi: 10.1109/TITS.2022.3153815.

[26] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets, and challenges," *IEEE Intell. Transp. Syst. Mag.*, vol. 15, no. 6, pp. 131–151, Nov. 2023, doi: 10.1109/MITS.2023.3298534.

[27] S. Liu et al., "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," 2023, *arXiv:2308.16714*.

[28] T. Huang et al., "Vehicle-to-everything cooperative perception for autonomous driving," 2023, *arXiv:2310.03525*.

[29] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. 71, Mar. 2021, doi: 10.1136/bmj.n71.

[30] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Inf. Softw. Technol.*, vol. 55, no. 12, pp. 2049–2075, Dec. 2013, doi: 10.1016/j.infsof.2013.07.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584913001560

[31] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499, doi: 10.1109/CVPR.2018.00472.

[32] J. Guo et al., "CoFF: Cooperative spatial feature fusion for 3-D object detection on autonomous vehicles," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11078–11087, Jul. 2021, doi: 10.1109/JIOT.2021.3053184.

[33] J. Guo et al., "Slim-FCP: Lightweight-feature-based cooperative perception for connected automated vehicles," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 15630–15638, Sep. 2022.

[34] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "V2VFormer: Vehicle-to-vehicle cooperative perception with spatial-channel transformer," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 2, pp. 3384–3395, Feb. 2024, doi: 10.1109/TIV.2024.3353254.

[35] Q. Xie, X. Zhou, T. Qiu, Q. Zhang, and W. Qu, "Soft actor–critic-based multilevel cooperative perception for connected autonomous vehicles," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21370–21381, Nov. 2022, doi: 10.1109/JIOT.2022.3179739.

[36] C. Liu, Y. Chen, J. Chen, R. Payton, M. Riley, and S.-H. Yang, "Cooperative perception with learning-based V2V communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 11, pp. 1831–1835, Nov. 2023, doi: 10.1109/LWC.2023.3295612.

[37] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "Dynamic feature sharing for cooperative perception from point clouds," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 3970–3976, doi: 10.1109/ITSC57777.2023.10422242.

[38] S. Teufel, J. Gamerdinger, G. Volk, and O. Bringmann, "Collective PV-RCNN: A novel fusion technique using collective detections for enhanced local LiDAR-based perception," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 1828–1834, doi: 10.1109/ITSC57777.2023.10422079.

[39] H. Chen, H. Wang, Z. Liu, D. Gu, and W. Ye, "HP3D-V2V: High-precision 3D object detection vehicle-to-vehicle cooperative perception algorithm," *Sensors*, vol. 24, no. 7, p. 2170, Mar. 2024, doi: 10.3390/s24072170.

[40] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12689–12697, doi: 10.1109/CVPR.2019.01298.

[41] Y. Ma et al., "MACP: Efficient model adaptation for cooperative perception," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 3361–3370, doi: 10.1109/WACV57701.2024.00334.

[42] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[43] G. Luo, H. Zhang, Q. Yuan, and J. Li, "Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3578–3586, doi: 10.1145/3503161.3548197.

[44] D. Qiao and F. Zulkernine, "Adaptive feature fusion for cooperative perception using LiDAR point clouds," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1186–1195, doi: 10.1109/WACV56688.2023.00124.

[45] Y. Li, S. Ren, P. Wu, S. Chen, F. Chen, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 1–8. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/f702defbc67edb455949f46babab0c18-Paper.pdf

[46] S. Ren, Z. Lei, Z. Wang, S. Chen, and W. Zhang, "Robust collaborative perception against communication interruption," in *Proc. Artif. Intell. Auton. Driving-Learn-Race Learn-Race (IJCAI-ECAI Workshop + Challenge)*, May 2022, pp. 1–8. [Online]. Available: https://learn-to-race.org/workshop-ai4ad-ijcai2022/assets/papers/paper_9.pdf

[47] J. Wang et al., "F-transformer: Point cloud fusion transformer for cooperative 3D object detection," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2022, pp. 171–182, doi: 10.1007/978-3-031-15919-0_15.

[48] S. Wei et al., "Asynchrony-robust collaborative perception via bird's eye view flow," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 28462–28477. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/5a829e299ebc1c1615ddb09e98fb6ce8-Paper-Conference.pdf

[49] 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 *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Macau, China, Oct. 2022, pp. 1743–1749, doi: 10.1109/ITSC55140.2022.9921947.

[50] J. Wang, Y. Zeng, and Y. Gong, "Collaborative 3D object detection for autonomous vehicles via learnable communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9804–9816, Sep. 2023, doi: 10.1109/TITS.2023.3272027.

[51] J. Gu et al., "FeaCo: Reaching robust feature-level consensus in noisy pose conditions," in *Proc. 31st ACM Int. Conf. Multimedia*. Ottawa, ON, Canada: ACM, Oct. 2023, pp. 3628–3636, doi: 10.1145/3581783.3611880.

[52] D. Yang et al., "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Nov. 2023, pp. 28462–28477. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/4f31327e046913c7238d5b671f5d820e-Paper-Conference.pdf

[53] K. Yang, D. Yang, J. Zhang, H. Wang, P. Sun, and L. Song, "What2comm: Towards communication-efficient collaborative perception via feature decoupling," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7686–7695, doi: 10.1145/3581783.3611699.

[54] J. Li et al., "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 4, pp. 2650–2660, Apr. 2023, doi: 10.1109/TIV.2023.3260040.

[55] Y. Lu et al., "Robust collaborative 3D object detection in presence of pose errors," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4812–4818, doi: 10.1109/ICRA48891.2023.10160546.

[56] K. Yang et al., "Spatio-temporal domain awareness for multi-agent collaborative perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23326–23335, doi: 10.1109/ICCV51070.2023.02137.

[57] X. Kong, W. Jiang, J. Jia, Y. Shi, R. Xu, and S. Liu, "DUSA: Decoupled unsupervised Sim2Real adaptation for vehicle-to-everything collaborative perception," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1943–1954, doi: 10.1145/3581783.3611948.

[58] Y. Liu, B. Sun, Y. Li, Y. Hu, and F.-Y. Wang, "HPL-ViT: A unified perception framework for heterogeneous parallel LiDARs in V2V," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 16417–16424, doi: 10.1109/ICRA57147.2024.10611513.

[59] L. Zhou, Z. Gan, and J. Fan, "CenterCoop: Center-based feature aggregation for communication-efficient vehicle-infrastructure cooperative 3D object detection," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3570–3577, Apr. 2024, doi: 10.1109/LRA.2023.3339399.

[60] M.-Q. Dao, J. S. Berrio, V. Frémont, M. Shan, E. Héry, and S. Worrall, "Practical collaborative perception: A framework for asynchronous and multi-agent 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 1–13, Sep. 2024, doi: 10.1109/TITS.2024.3371177.

[61] Z. Lei et al., "Robust collaborative perception without external localization and clock devices," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 7280–7286, doi: 10.1109/ICRA57147.2024.10610635.

[62] J. Li et al., "S2R-ViT for multi-agent cooperative perception: Bridging the gap from simulation to reality," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 16374–16380.

[63] Y. Liu et al., "Select2Col: Leveraging spatial–temporal importance of semantic information for efficient collaborative perception," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12556–12569, Sep. 2024, doi: 10.1109/TVT.2024.3390414.

[64] P. Liu, Z. Wang, G. Yu, B. Zhou, and P. Chen, "Region-based hybrid collaborative perception for connected autonomous vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 3119–3128, Mar. 2024, doi: 10.1109/TVT.2023.3324439.

[65] J. Wang, X. Guo, H. Wang, P. Jiang, T. Chen, and Z. Sun, "Pillar-based cooperative perception from point clouds for 6G-enabled cooperative autonomous vehicles," *Wireless Commun. Mobile Comput.*, vol. 2022, no. 1, pp. 1–13, Jul. 2022, doi: 10.1155/2022/3646272.

[66] Z. Li, H. Liang, H. Wang, M. Zhao, J. Wang, and X. Zheng, "MKD-cooper: Cooperative 3D object detection for autonomous driving via multi-teacher knowledge distillation," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1490–1500, Jan. 2024, doi: 10.1109/TIV.2023.3310580.

[67] L. Wang, J. Lan, and M. Li, "PAFNet: Pillar attention fusion network for vehicle–infrastructure cooperative target detection using LiDAR," *Symmetry*, vol. 16, no. 4, p. 401, Mar. 2024, doi: 10.3390/sym16040401.

[68] Z. Bai et al., "Pillar attention encoder for adaptive cooperative perception," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 24998–25009, Jul. 2024, doi: 10.1109/JIOT.2024.3390552.

[69] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake LiDAR in 3D detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9243–9252, doi: 10.1109/CVPR52729.2023.00892.

[70] S. Huang, J. Zhang, Y. Li, and C. Feng, "ActFormer: Scalable collaborative perception via active queries," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 14716–14723, doi: 10.1109/ICRA57147.2024.10610997.

[71] Z. Wang et al., "EMIFF: Enhanced multi-scale image feature fusion for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 16388–16394, doi: 10.1109/ICRA57147.2024.10610545.

[72] S. Fan, H. Yu, W. Yang, J. Yuan, and Z. Nie, "QUEST: Query stream for practical cooperative perception," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 18436–18442, doi: 10.1109/ICRA57147.2024.10610214.

[73] Y. Lee, J.-W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 752–760, doi: 10.1109/CVPRW.2019.00103.

[74] H. Yin et al., "V2VFormer++: Multi-modal vehicle-to-vehicle cooperative perception via global-local transformer," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 2153–2166, Feb. 2024, doi: 10.1109/TITS.2023.3314919.

[75] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. 36th Adv. Neural Inf. Process. Syst. (NeurIPS)*, Sep. 2022, pp. 1–12. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/1f5c5cd01b864d53cc5fa0a3472e152e-Paper-Conference.pdf

[76] F. Chi, Y. Wang, M. T. Pourazad, P. Nasiopoulos, and V. C. M. Leung, "Multimodal cooperative 3D object detection over connected vehicles for autonomous driving," *IEEE Netw.*, vol. 37, no. 4, pp. 265–272, Jul. 2023, doi: 10.1109/MNET.010.2300029.

[77] Y. Lu, Y. Hu, Y. Zhong, D. Wang, S. Chen, and Y. Wang, "An extensible framework for open heterogeneous collaborative perception," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, May 2024, pp. 1–7. [Online]. Available: https://iclr.cc/virtual/2024/poster/18889

[78] Y. Zhou, C. Yang, P. Wang, C. Wang, X. Wang, and N. Ngoc Van, "ViT-FuseNet: Multimodal fusion of vision transformer for vehicle-infrastructure cooperative perception," *IEEE Access*, vol. 12, pp. 31640–31651, 2024, doi: 10.1109/ACCESS.2024.3368404.

[79] L. Zhang, B. Wang, Z. Wang, and Y. Zhao, "V2VFusion: Multimodal fusion for enhanced vehicle-to-vehicle cooperative perception," in *Proc. China Autom. Congr. (CAC)*, Nov. 2023, pp. 3691–3696, doi: 10.1109/CAC59555.2023.10450676.

[80] H. Zhang, G. Luo, Y. Cao, Y. Jin, and Y. Li, "Multi-modal virtual-real fusion based transformer for collaborative perception," in *Proc. IEEE 13th Int. Symp. Parallel Archit., Algorithms Program. (PAAP)*, Nov. 2022, pp. 1–6, doi: 10.1109/PAAP56126.2022.10010640.

[81] X. Kuang, H. Zhu, B. Yu, and B. Li, "Fast clustering for cooperative perception based on LiDAR adaptive dynamic grid encoding," *Cognit. Comput.*, vol. 16, no. 2, pp. 546–565, Mar. 2024, doi: 10.1007/s12559-023-10211-x.

[82] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 30, Dec. 2022, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[83] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.

[84] 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 *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 605–621, doi: 10.1007/978-3-030-58536-5_36.

[85] R. Xu, W. Chen, H. Xiang, X. Xia, L. Liu, and J. Ma, "Model-agnostic multi-agent perception framework," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 1471–1478, doi: 10.1109/ICRA48891.2023.10161460.

[86] H. Yu, Y. Zhao, Y. Zou, Q. Li, H. Yu, and Y. Ren, "Multistage fusion approach of LiDAR and camera for vehicle-infrastructure cooperative object detection," in *Proc. 5th World Conf. Mech. Eng. Intell. Manuf. (WCMEIM)*, Nov. 2022, pp. 811–816, doi: 10.1109/WCMEIM56910.2022.10021459.

[87] Y. Yuan, H. Cheng, and M. Sester, "Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3054–3061, Apr. 2022, doi: 10.1109/LRA.2022.3143299.

[88] A. Ben Khalifa, I. Alouani, M. A. Mahjoub, and A. Rivenq, "A novel multi-view pedestrian detection database for collaborative intelligent transportation systems," *Future Gener. Comput. Syst.*, vol. 113, pp. 506–527, Dec. 2020, doi: 10.1016/j.future.2020.07.025.

[89] D. Marez, L. Nans, and S. Borden, "Bandwidth constrained cooperative object detection in images," *Proc. SPIE*, vol. 12276, pp. 128–140, Oct. 2022, doi: 10.1117/12.2636279.

[90] R. Mao et al., "MoRFF: Multi-view object detection for connected autonomous driving under communication and localization limitations," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, Oct. 2023, pp. 1–7, doi: 10.1109/VTC2023-FALL60731.2023.10333428.

[91] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6876–6883, doi: 10.1109/ICRA40945.2020.9197364.

[92] N. Glaser, Y.-C. Liu, J. Tian, and Z. Kira, "Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 2406–2413, doi: 10.1109/IROS51168.2021.9636761.

[93] Y. Yuan, H. Cheng, M. Y. Yang, and M. Sester, "Generating evidential BEV maps in continuous driving space," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 27–41, Oct. 2023, doi: 10.1016/j.isprsjprs.2023.08.013.

[94] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," in *Proc. Conf. Robot Learn.*, 2023, pp. 989–1000. [Online]. Available: https://proceedings.mlr.press/v205/xu23a/xu23a.pdf

[95] H. Liu, Z. Gu, C. Wang, P. Wang, and D. Vukobratovic, "A LiDAR semantic segmentation framework for the cooperative vehicle-infrastructure system," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, Oct. 2023, pp. 1–5, doi: 10.1109/VTC2023-FALL60731.2023.10333790.

[96] R. Song et al., "Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17996–18006, doi: 10.1109/CVPR52733.2024.01704.

[97] H. Su, S. Arakawa, and M. Murata, "3D multi-object tracking based on two-stage data association for collaborative perception scenarios," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–7, doi: 10.1109/IV55152.2023.10186777.

[98] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D LiDAR," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 8, pp. 4069–4080, Aug. 2023, doi: 10.1109/TIV.2023.3282567.

[99] H. Yu et al., "V2X-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5486–5495, doi: 10.1109/CVPR52729.2023.00531.

[100] S. Su et al., "Collaborative multi-object tracking with conformal uncertainty propagation," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3323–3330, Apr. 2024, doi: 10.1109/LRA.2024.3364450.

[101] H.-K. Chiu, C.-Y. Wang, M.-H. Chen, and S. F. Smith, "Probabilistic 3D multi-object cooperative tracking for autonomous driving via differentiable multi-sensor Kalman filter," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 18458–18464, doi: 10.1109/ICRA57147.2024.10610487.

[102] C. Chang et al., "BEV-V2X: Cooperative birds-eye-view fusion and grid occupancy prediction via V2X-based data sharing," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 11, pp. 4498–4514, Nov. 2023, doi: 10.1109/TIV.2023.3293954.

[103] T. Wang et al., "DeepAccident: A motion and accident prediction benchmark for V2X autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 6, pp. 5599–5606, doi: 10.1609/aaai.v38i6.28370.

[104] X. Ding, X. Zhang, J. Han, and G. Ding, "Diverse branch block: Building a convolution as an inception-like unit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10886–10895, doi: 10.1109/CVPR46437.2021.01074.

[105] Y. Zhang et al., "BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," 2022, *arXiv:2205.09743*.

[106] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[107] A. H. Sakr, "Cooperative road geometry estimation via sharing processed camera data," in *Proc. IEEE 3rd Connected Automated Vehicles Symp. (CAVS)*, Nov. 2020, pp. 1–6, doi: 10.1109/CAVS51000.2020.9334579.

[108] J. Gamerdinger, S. Teufel, G. Volk, and O. Bringmann, "CoLD fusion: A real-time capable spline-based fusion algorithm for collective lane detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–8, doi: 10.1109/IV55152.2023.10186632.

[109] L. L. F. Jahn, S. Park, Y. Lim, J. An, and G. Choi, "Enhancing lane detection with a lightweight collaborative late fusion model," *Robot. Auto. Syst.*, vol. 175, May 2024, Art. no. 104680, doi: 10.1016/j.robot.2024.104680.

[110] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng, "Multi-robot scene completion: Towards task-agnostic collaborative perception," in *Proc. 6th Conf. Robot Learn. (CoRL)*, 2022, pp. 2062–2072. [Online]. Available: https://proceedings.mlr.press/v205/li23e/li23e.pdf

[111] B. Wang, L. Zhang, Z. Wang, Y. Zhao, and T. Zhou, "Core: Cooperative reconstruction for multi-agent perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8676–8686, doi: 10.1109/ICCV51070.2023.00800.

[112] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, doi: 10.1145/358669.358692.

[113] C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: A survey," *Comput. Vis. Image Understand.*, vol. 117, no. 11, pp. 1610–1627, Nov. 2013, doi: 10.1016/j.cviu.2013.07.004.

[114] T. Wang et al., "UMC: A unified bandwidth-efficient and multi-resolution based collaborative perception framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8153–8162, doi: 10.1109/ICCV51070.2023.00752.

[115] H. Yu, Y. Tang, E. Xie, J. Mao, P. Luo, and Z. Nie, "Flow-based feature fusion for vehicle-infrastructure cooperative 3D object detection," in *Proc. 37th Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2023, pp. 1–14. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/6ca5d2665de83394f437dad0c3746907-Paper-Conference.pdf

[116] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:1257163

[117] A. Bazzi, B. M. Masini, A. Zanella, and I. Thibault, "On the performance of IEEE 802.11p and LTE-V2V for the cooperative awareness of connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10419–10432, Nov. 2017.

[118] M. Gonzalez-Martín, M. Sepulcre, R. Molina-Masegosa, and J. Gozalvez, "Analytical models of the performance of C-V2X mode 4 vehicular communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1155–1166, Feb. 2019.

[119] Y. Sheng, H. Ye, L. Liang, S. Jin, and G. Y. Li, "Semantic communication for cooperative perception based on importance map," *J. Franklin Inst.*, vol. 361, no. 6, Apr. 2024, Art. no. 106739, doi: 10.1016/j.jfranklin.2024.106739.

[120] J. Li, B. Li, X. Liu, R. Xu, J. Ma, and H. Yu, "Breaking data silos: Cross-domain learning for multi-agent perception from independent private sources," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 18414–18420, doi: 10.1109/ICRA57147.2024.10610591.

[121] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 6035–6042, doi: 10.1109/ICRA48891.2023.10160871.

[122] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628, doi: 10.1109/CVPR42600.2020.01164.

[123] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2446–2454, doi: 10.1109/CVPR42600.2020.00252.

[124] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21329–21338, doi: 10.1109/CVPR52688.2022.02067.

[125] C. Ma et al., "HoloVic: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 22129–22138, doi: 10.1109/CVPR52733.2024.02089.

[126] J. Axmann et al., "LUCOOP: Leibniz university cooperative perception and urban navigation dataset," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–8, doi: 10.1109/IV55152.2023.10186693.

[127] R. Xu et al., "V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13712–13722, doi: 10.1109/CVPR52729.2023.01318.

[128] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "TUMTraf V2X cooperative perception dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 22668–22677, doi: 10.1109/CVPR52733.2024.02139.

[129] 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 *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 1155–1162, doi: 10.1109/ITSC48978.2021.9564825.

[130] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2583–2589, doi: 10.1109/ICRA46639.2022.9812038.

[131] Y. Li et al., "V2X-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022, doi: 10.1109/LRA.2022.3192802.

[132] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "VINet: Lightweight, scalable, and heterogeneous cooperative perception for 3D object detection," *Mech. Syst. Signal Process.*, vol. 204, Dec. 2023, Art. no. 110723, doi: 10.1016/j.ymssp.2023.110723.

[133] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.

[134] X. Cai et al., "Analyzing infrastructure LiDAR placement with realistic LiDAR simulation library," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 5581–5587, doi: 10.1109/ICRA48891.2023.10161027.

[135] C. Liu, J. Chen, Y. Chen, R. Payton, M. Riley, and S.-H. Yang, "Self-supervised adaptive weighting for cooperative perception in V2V communications," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 2, pp. 3569–3580, Feb. 2024, doi: 10.1109/TIV.2023.3345035.

[136] S. Su et al., "Uncertainty quantification of collaborative detection for self-driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 5588–5594, doi: 10.1109/ICRA48891.2023.10160367.

[137] S. Shi, C. Zhang, A. Lv, and S. He, "MCoT: Multi-modal vehicle-to-vehicle cooperative perception with transformers," in *Proc. IEEE 29th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2023, pp. 1612–1619, doi: 10.1109/ICPADS60453.2023.00226.

Lei Wan received the bachelor's degree in mechanical engineering from Harbin Engineering University, China, and the master's degree in mechatronics and information technology from Karlsruhe Institute of Technology (KIT), Germany, where he is currently pursuing the Ph.D. degree. He is a Full-Time Research Engineer with XITASO GmbH, Germany. His research interests include autonomous driving, computer vision, sensor fusion, and collaborative perception with V2X, whose goal is to improve the safety of autonomous driving through the development of robust perception systems.

Jianxin Zhao is currently a Post-Doctoral Researcher with Karlsruhe Institute of Technology (KIT), Germany. His research interests include scientific computation, machine learning, and their application and optimization in the real world. In the post-doctoral work, he is researching the use of machine learning methods in cooperative perception, with a focus on the impact of high dynamism in the cooperative autonomous systems.

Andreas Wiedholz received the bachelor's degree in computer engineering and the master's degree in robotics and AI from the Technical University of Applied Sciences Augsburg. He is currently a Full-Time Researcher with XITASO GmbH, Germany. His research interests include perception and modeling of autonomous behavior in robotic systems.

Manuel Bied received the bachelor's and master's degree in electrical engineering and information technology from Technical University Darmstadt, Germany, and the Ph.D. degree in robotics from Sorbonne University, France. He is currently a Post-Doctoral Researcher with Karlsruhe Institute of Technology (KIT), Germany. In his work, he is focusing on the use of robots in traffic with the aim of increasing road safety. His research interests include human–robot interaction, V2X, machine learning, and collective perception.

Mateus Martinez de Lucena received the degree in computer engineering from the Federal University of Amazonas, and the master's degree in computer science from UFSC. He is currently pursuing the Ph.D. degree with the Federal University of Santa Catarina, supervised by Prof. Dr. Antônio Augusto Fröhlich. He is researching collaborative perception in vehicular networks from a consensus perspective.

Abhishek Dinkar Jagtap is currently pursuing the Ph.D. degree with the Technische Hochschule Ingolstadt, Germany, and affiliated with the Center of Automotive Research on Integrated Safety Systems and Measurement Area (CARISSMA), a leading research and testing facility for vehicle safety. He received the master's degree in robotics and autonomous systems from the Universität zu Lübeck, Germany. His research interests include vehicle-to-everything (V2X) communication, computer vision, and cooperative perception. Specifically, he aims to enhance the efficiency of V2X communication-based cooperative perception by leveraging intermediate features from deep learning models.

Andreas Festag (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Berlin Institute of Technology, Berlin, Germany, in 2003. He has held research positions with the Telecommunication Networks Group, Berlin Institute of Technology, the Heinrich-Hertz-Institute (HHI), Berlin, NEC Laboratories, Heidelberg, Germany, the Vodafone Chair Mobile Communication Systems, Dresden University of Technology, Dresden, Germany, and the Fraunhofer Institute for Transportation and Infrastructure Systems, Dresden. He is currently a Professor with the Technische Hochschule Ingolstadt, Germany, and is affiliated with the Center of Automotive Research on Integrated Safety Systems and Measurement Area (CARISSMA). He is also the Deputy Head of the Fraunhofer Application Center for "Connected Mobility and Infrastructure, "Ingolstadt. His research interests include architecture, design, and performance evaluation of wireless and mobile communication systems and protocols, with an emphasis on vehicular communication and cooperative intelligent transportation systems (C-ITS).

Antônio Augusto Fröhlich (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Technical University of Berlin, Germany. He has coordinated several research and development projects on embedded systems, including the ALTATV Open, Free, Scalable Digital TV Platform, the CIA^2 research network on Smart Cities and the Internet of Things, and the Smart Campus project at UFSC. Significant contributions from these projects materialized within the Brazilian Digital Television System (SBTVD) and IoT technology for energy distribution, smart cities, and precision agriculture. He is currently a Full Professor with the Federal University of Santa Catarina (UFSC), Brazil, where he has been leading the Software/Hardware Integration Laboratory (LISHA), since 2001. He is a senior member of ACM and SBC.

Hannan Ejaz Keen received the B.Sc. degree in electrical engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, the M.S. degree in electrical engineering from Lahore University of Management Sciences (LUMS), Lahore, and the Ph.D. degree in autonomous off-road robotics from Robotics Research Laboratory, RPTU Kaiserslautern-Landau, Germany. He was a Researcher with the Robotics Research Laboratory. He is currently a Senior Researcher with the Team of Autonomous Systems, XITASO GmbH. His research interests include sensor fusion, perception, and mapping.

Alexey Vinel (Senior Member, IEEE) received the Ph.D. degree from the Tampere University of Technology, Finland, in 2013. He was a Professor with the University of Passau, Germany. Since 2015, he has been a Professor with Halmstad University, Sweden (now part-time). He is currently a Professor with Karlsruhe Institute of Technology (KIT), Germany. His research interests include vehicular communications and networking, cooperative automated and autonomous driving, and future smart mobility solutions.