The Components of Collaborative Joint Perception and Prediction: A Conceptual Framework

Lei Wan[1,2][a], Hannan Ejaz Keen[1][b] and Alexey Vinel[2][c]
[1]*XITASO GmbH, Augsburg, Germany*
[2]*Karlsruhe Institut of Technology, Karlsruhe, Germany*
{*lei.wan, hannan.keen*}*@xitaso.com, alexey.vinel@kit.edu*

Keywords: Autonomous Driving, Connected Autonomous Vehicles, Cooperative-Intelligent Transportation Systems, Collaborative Perception, Collaborative Joint Perception and Prediction.

Abstract: Connected Autonomous Vehicles (CAVs) benefit from Vehicle-to-Everything (V2X) communication, which enables the exchange of sensor data to achieve Collaborative Perception (CP). To reduce cumulative errors in perception modules and mitigate the visual occlusion, this paper introduces a new task, Collaborative Joint Perception and Prediction (Co-P&P), and provides a conceptual framework for its implementation to improve motion prediction of surrounding objects, thereby enhancing vehicle awareness in complex traffic scenarios. The framework consists of two decoupled core modules, Collaborative Scene Completion (CSC) and Joint Perception and Prediction (P&P) module, which simplify practical deployment and enhance scalability. Additionally, we outline the challenges in Co-P&P and discuss future directions for this research area.

1 INTRODUCTION

Autonomous Driving (AD) technology is essential for advancing intelligent transportation systems, contributing to improved road safety, enhanced traffic efficiency, energy conservation, and reduced carbon emissions. A key component of the AD framework is perception, which involves detecting dynamic objects and interpreting the static environment. The perception module encompasses various tasks, including object detection, tracking, motion prediction, and semantic segmentation. Traditionally, these tasks are implemented in a modular format, forming the basis for downstream functions like planning and control (Keen and Berns, 2023; Keen and Berns, 2020). Advancements in artificial intelligence and sensor fusion have significantly improved vehicle perception capabilities. However, single-vehicle perception still faces challenges, particularly with visual occlusion, which can pose safety risks and lead to accidents. Vehicle-to-Everything (V2X) technology offers a promising approach to address these limitations by enabling the sharing data with other vehicles or infrastructure, effectively enhancing perception and mitigating occlusion issues.

With V2X communication, Connected Autonomous Vehicles (CAVs) can achieve Collaborative Perception (CP) by integrating data from multiple sources. Initial research on CP began within the communication field, focusing on standardizing V2X message types and optimizing communication efficiency. Recently, CP research has expanded into computer vision and robotics, where the emphasis is shifting from sharing standardized messages, such as Cooperative Perception Message (CPM) containing detected objects, to share raw sensor data or neural features. For example, Chen et al. (Chen et al., 2019) propose a feature-based CP approach that transmits and combines LiDAR features across vehicles, enhancing perception performance within bandwidth constraints. Similarly, Hu et al. (Hu et al., 2023) present a camera-based CP method that integrates visual Bird's Eye View (BEV) features from multiple agents, providing a more comprehensive view of dynamic objects.

The use of collaborative methodologies extend beyond object detection to enhance other perception tasks. For instance, Liu et al. (Liu et al., 2023) introduce a collaborative semantic segmentation framework utilizing intermediate collaboration, achieving superior results compared to single-vehicle methods. In motion prediction, Wang et al. (Wang et al., 2020)

[a] https://orcid.org/0009-0007-4470-9088
[b] https://orcid.org/0009-0001-6217-9427
[c] https://orcid.org/0000-0003-4894-4134

Wan, L., Keen, H. E. and Vinel, A.
The Components of Collaborative Joint Perception and Prediction: A Conceptual Framework.
DOI: 10.5220/0013285300003941
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2025)*, pages 458-465
ISBN: 978-989-758-745-0; ISSN: 2184-495X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

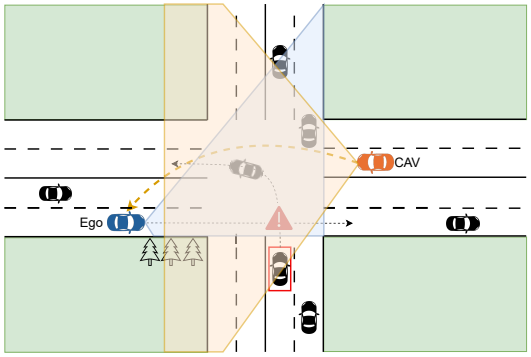

Figure 1: Schematic diagram of Collaborative Perception. The diagram illustrates a scenario at an intersection where two CAVs collaborate to enhance perception. The ego vehicle (blue) has a limited field of view due to occlusions, such as trees and buildings, which block its line of sight to a vehicle turning left. A collaborating vehicle (orange) positioned across the intersection shares its sensor data, expanding the ego vehicle's awareness. The shaded areas represent the Field of View (FOV) for each vehicle.

demonstrate how collaboration enhances the precision of predicted trajectories.Additionally some perception tasks are handled within multi-task pipelines, such as the V2XFormer model by Wang et al. (Wang et al., 2024), which simultaneously outputs detection, motion prediction, and semantic segmentation results.While multi-task approaches benefit from resource savings by sharing a common backbone, they often overlook temporal information across sensor frames, which is essential for tracking and motion prediction.

An emerging trend is the development of differentiable frameworks that seamlessly integrate various perception tasks within a unified model, enabling end-to-end training. For example, Liang et al. (Liang et al., 2020) propose an end-to-end Joint Perception and Prediction (P&P) framework for single vehicles equipped with LiDAR. Similarly, Gu (Gu et al., 2023) introduces a camera-only end-to-end pipeline for P&P, utilizing visual features to achieve both detection and motion prediction. These approaches highlight the potential of end-to-end learning to address bottlenecks in traditional perception pipelines, where cumulative errors across stages can degrade performance. With end-to-end learning, cumulative noise is mitigated, and motion prediction benefits significantly from the integration of fine-grained contextual information. Nonetheless, current research in P&P still encounters challenges, particularly with visual occlusion, which significantly impacts prediction accuracy for obscured targets.

To overcome this issue, we propose the Collaborative Joint Perception and Prediction (Co-P&P) framework, which incorporates V2X collaboration. Our framework is based on the premise that CP complements ego-vehicle perception, making it adaptable to scenarios with or without V2X support. Additionally, to simplify deployment and enhance scalability, our framework decouples the training of the collaboration module from perception tasks. Inspired by recent work (Li et al., 2022; Wang et al., 2023a), our approach uses collaborative scene completion to address visual occlusion. Consequently, the Co-P&P framework comprises two core modules: the Collaborative Scene Completion (CSC) module and the Joint Perception and Prediction (P&P) module.

In addition to developing CP approaches, establishing effective evaluation methodologies is crucial for advancing CP research. Current studies largely adopt evaluation methods designed for single-vehicle perception. Notably, only one study (Wang et al., 2023b) has introduced an evaluation focused on invisible objects, highlighting CP's potential to address visual occlusion. New evaluation methods are also required to assess the performance of Co-P&P.

The main contributions of this paper are as follows:

- We introduce a conceptual framework for Co-P&P, designed to address cumulative errors inherent in modular designs and mitigate visual occlusion challenges.
- We present a re-formulation of evaluation methods in CP and propose an evaluation approach tailored for Co-P&P that aligns with the motivation of V2X collaboration.
- We outline the challenges surrounding Co-P&P, and suggest future work to further enhance this framework.

The structure of the paper is as follows: Section 2 details the system design, while Section 3 introduces the evaluation method for Co-P&P. Sections 4 discusses real-world challenges regarding practical deployment. Finally, Section 5 concludes with a summary and outlook.

2 DETAILS OF THE SYSTEM

Figure 2 presents the conceptual framework, encompassing sensors, localization, High Definition Map (HD Map), communication, P&P, collaborative scene completion. This section provides an overview of each core component of the framework along with their corresponding approaches.

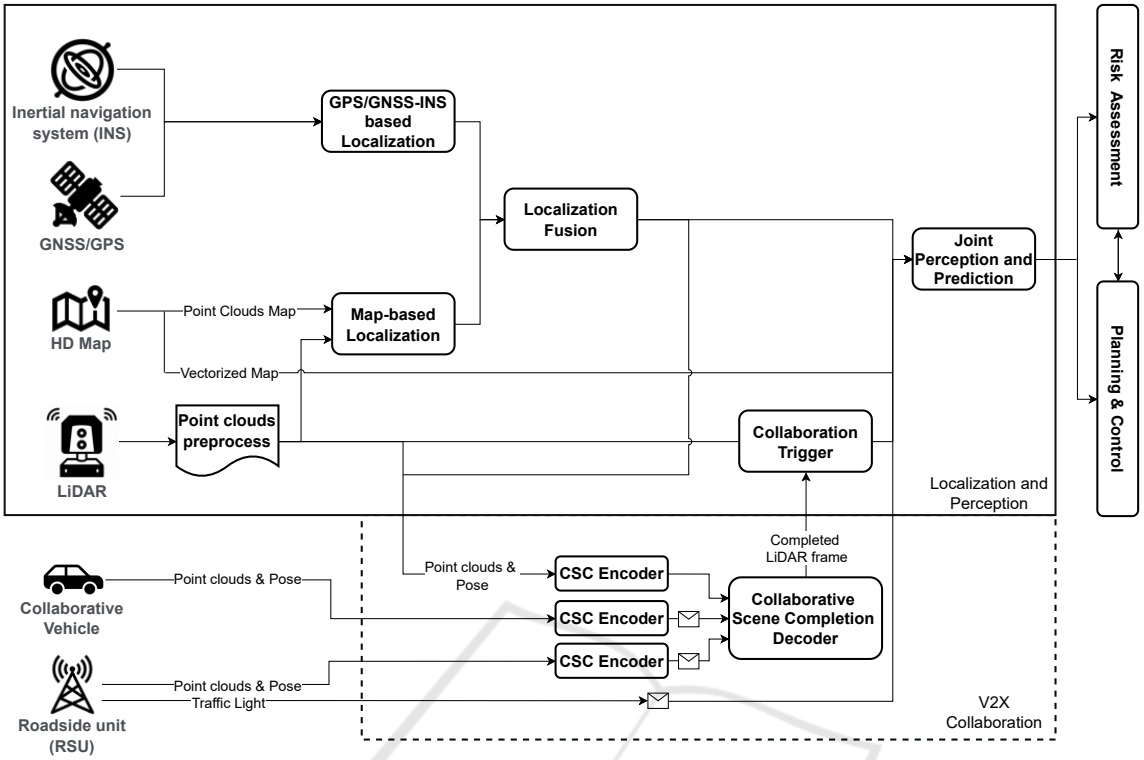

Figure 2: Schematic diagram of Collaborative Joint Perception and Prediction.The system combines GPS/GNSS-INS and map-based localization for precise positioning. Sensing data (point clouds, poses) from collaborative vehicle and roadside unit are processed and shared with the ego vehicle, enabling Collaborative Scene Completion (CSC) to provide a comprehensive LiDAR frame. To optimize bandwidth usage, intermediate features are shared for scene completion instead of raw point clouds are shared via V2X. The collaboration trigger manages CSC activation. The P&P module integrates localization, map, and LiDAR data to jointly enhance perception and prediction, which feeds into the risk assessment and planning and control modules for real-time decision-making.

2.1 Sensors

To capture a 3D view of the environment, various sensor types can be employed, including LiDAR, radar, and different types of cameras such as RGB and infrared camera. In our conceptual framework, LiDAR serves as the primary sensor due to its high precision in 3D measurements, significantly enhancing the 3D perception capabilities of Autonomous Vehicles (AVs).

LiDAR sensors vary in their scanning patterns, typically classified into spinning and oscillating types (Triess et al., 2021). Spinning LiDAR uses a regular scanning pattern that provides an even distribution of points across a 360° FOV. In contrast, oscillating LiDAR follows a snake-like pattern, creating a denser yet uneven point distribution within a constrained FOV. Each type offers distinct characteristics that can lead to a domain gap in perception models due to differences in data representation. Addressing this domain gap across different LiDAR sensor is essential for CP systems.

2.2 Localization

Beyond environmental perception, precise self-localization is essential for CP. Accurate localization allows for data fusion across dynamic agents by establishing a consistent coordinate system to align all sensory data. Thus, the effectiveness of CP depends significantly on the localization accuracy of CAVs.

Traditionally, vehicles rely on Global Navigation Satellite System (GNSS) or Global Positioning System (GPS) to determine their position using trilateration. However, GNSS-based methods face challenges like Non-Line-of-Sight and multipath propagation, often resulting in errors exceeding 3 meters, which undermines reliability and safety in AD (Ochieng and Sauer, 2002). HD Map can mitigate localization errors, achieving centimeter-level accuracy (Chalvatzaras et al., 2022). These maps are created through extensive data collection runs, often using LiDAR to construct a detailed point cloud layer. For precise vehicle positioning on HD Map, both GNSS and LiDAR sensors are used, providing a precise po-

sitioning approach.

2.3 HD Map

HD Map serves not only for localization but also offer essential semantic information about the static environment. They include detailed road data such as lane boundaries, lane centerlines, road markings, traffic signs, poles, and traffic light locations. This information aids vehicles in interpreting traffic rules and understanding the surrounding environment, enhancing motion prediction accuracy. Xu et al. (Xu et al., 2023) highlight the significant impact of map quality on motion prediction performance, showing that high-quality, curated HD Map outperform systems relying on online mapping or operating without maps. In our framework, the map operates as an independent module that interfaces with the perception module. This design enables compatibility with various mapping solutions, supporting scalability to online mapping or even cost-efficient, mapless approaches.

2.4 Communication

V2X communication technology forms a critical foundation for CP. CAVs and intelligent infrastructure use sensors to perceive the environment and then transmit this data through V2X communication. Two primary technologies support V2X communications: Dedicated Short-Range Communication (DSRC) and cellular network technologies (Abboud et al., 2016).

DSRC is a wireless technology designed for automotive and Cooperative-Intelligent Transportation Systems (C-ITS) applications, allowing short-range information exchange between devices. It operates without additional network infrastructure and offers low latency, making it suitable for safety-critical applications (Kenney, 2011). However, DSRC has limitations, including a relatively short communication range and reduced scalability in scenarios with high vehicular density (Harding et al., 2014). Cellular networks, on the other hand, offer a potential solution for C-ITS by providing greater bandwidth. These capabilities ensure that sensor data, crucial for CP, can be effectively transmitted across distributed entities. Yet, some Cellular Vehicle-to-Everything (C-V2X) modes depend on cellular infrastructure, meaning performance may degrade in areas far from base stations, impacting latency.

Given the limitations of using either V2X technology alone, a hybrid approach that combines both DSRC and cellular technologies is more promising, enabling novel DSRC–cellular interworking schemes. In our framework, data such as traffic light information, which requires low bandwidth, is well-suited for DSRC. Meanwhile, sensor data, with its higher bandwidth demands, is more effectively handled by C-V2X.

2.5 Joint Perception and Prediction

The P&P module forms the core of our framework. This module integrates data from LiDAR, vehicle pose, HD Map, and traffic light information to generate detection results and forecast the trajectories of relevant agents, as depicted in Figure 3. The P&P pipeline includes a LiDAR encoder, temporal encoder, map encoder, multi-agent interaction encoder, and P&P decoder.

- **Lidar Encoder:** To enable semantic understanding of the surrounding environment for AVs, the LiDAR encoder extracts semantic features from point clouds. For example, VoxelNet (Zhou and Tuzel, 2018) divides the point cloud into a 3D voxel grid, aggregates features within each voxel, and encodes these features. By applying voxel convolution, it captures 3D spatial features, which are then flattened along the z-axis and transformed into BEV features, enhanceing computational efficiency
- **Spatial-Temporal Attention:** In addition to spatial features from the LiDAR encoder, temporal information across multiple frames is crucial for understanding temporal dynamics of the environment (Bharilya and Kumar, 2023). In our framework, the temporal encoder captures this temporal context from LiDAR BEV features. For instance, a cross-attention mechanism (Vaswani et al., 2017) extracts context between frames, generating spatial-temporal features.
- **Map Encoder:** Map and traffic light information are crucial for motion prediction (Ettinger et al., 2021). To interact with spatial-temporal features, the map and traffic light data are encoded as neural features. For instance, the HD Map is transformed into the ego-vehicle's coordinate system, centered on the ego-vehicle, and only map information within a defined surrounding area is used. Traffic light data are integrated into the map as environmental indicators and encoded as features.
- **Multi-Agent Interaction Attention:** Modeling interactions among multiple agents is challenging (Bharilya and Kumar, 2023). This module combines agent and map features to ensure more precise modeling of these interactions. This block first computes interactions between the map and

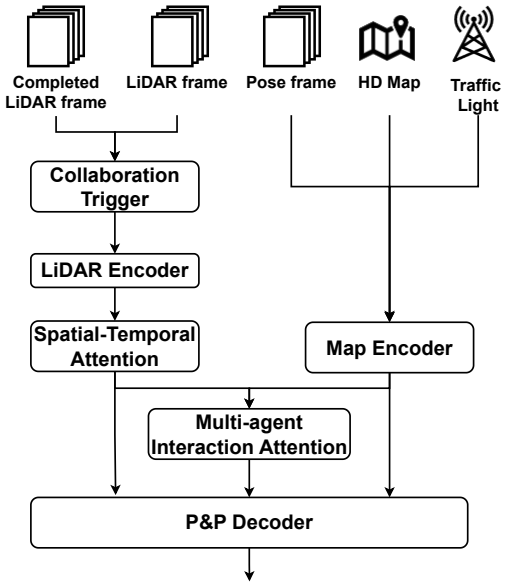

Figure 3: Schematic diagram of P&P.

agent information, then calculates inter-agent interactions within regions of interest (ROIs).

- **P&P Decoder:** The decoder integrates all relevant features to produce accurate perception and prediction outputs. Detection results are represented as a BEV map comprising a map mask and object masks. Motion prediction is represented as a BEV flow output, aligning well with downstream tasks such as planning and decision-making.

2.6 Collaboration Trigger

While multi-agent collaboration provides significant benefits for CAVs, it also demands substantial resources. Collaboration is often unnecessary when the ego vehicle has unobstructed visibility. To balance system effectiveness and efficiency, a collaboration trigger is needed to activate collaboration only at optimal times. Designing an effective collaboration trigger and identifying relevant decision factors remain underexplored areas of research (Huang et al., 2023). In our framework, we consider scenario occlusion levels, the confidence level of the ego vehicle's perception, and communication conditions in developing this trigger metric. When the metric value exceeds a specified threshold, the system activates the collaboration module

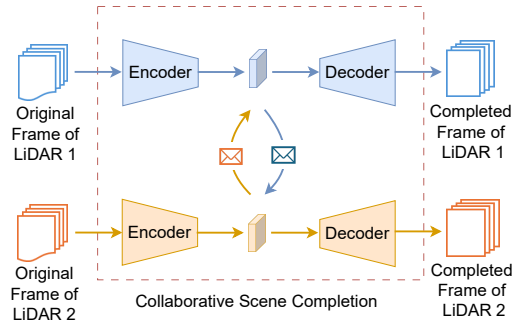

Figure 4: Schematic diagram of Collaborative Scene Completion (CSC).

2.7 Collaborative Scene Completion

Traditional CP frameworks generally involve the sharing of neural features generated by deep learning-based CP modules or the exchange of perception results among cooperative agents. However, this approach is task-specific, meaning that the shared data can only support particular perception tasks, leading to heterogeneity of systems that limits effective collaboration between diverse agents (Han et al., 2023). Additionally, conventional methods often require joint model training across agents and, in some cases, re-training the whole-model for each perception task (Li et al., 2022). This joint training can be impractical and resource-intensive in real-world applications.

In our framework, we decouple V2X collaboration from the P&P pipeline, allowing independent training of each component. V2X collaboration is managed through task-agnostic collaborative scene completion, which benefits all downstream tasks without needing task-specific data transmission. By reconstructing a comprehensive scene using latent features and sharing these features across agents, this approach minimizes communication demands, as shown in Figure 4. The completed scene is then fed into the P&P pipeline, as shown in Figure 3.

3 EVALUATION

In addition to developing the Co-P&P system,it is essential to establish effective evaluation methods to accurately access its performance. However, evaluating CP presents unique challenges. Most existing research relies on methods adapted from single-vehicle perception, which fail to reflect CP's capacity to address visual occlusions (Wang et al., 2023b). To overcome this limitation, new evaluation methods are required.

Table 1: Summary of Metrics for Evaluating Joint Perception and Prediction.

Metrics	Description
$minADE_k$	the minimum over k predictions of Average Distance Error: the average of point-wise L2 distances between the prediction and ground-truth forecasts
$minFDE_k$	the minimum over k predictions of Final Distance Error: the L2 distance at the final future time step
$MR_{k@x}$	the MissRate: the ratio of forecasts having $minFDE_k > x = 4m$
mAP_f	the Mean Forecasting Average Precision: adapted from mAP_{det}, mAP_f additionally penalizes trajectories that have correct first-frame detections but inaccurate forecasts ($minFDE_k < 4m$), but also trajectories with incorrect first-frame detections (center distance $< 2m$)

3.1 Evaluation Method

As CP complements single-vehicle perception, primarily aiming to resolve visual occlusions, evaluation metrics should reflect precision on objects that are hidden from an individual vehicle's view but visible from a collaborative perspective. For example, Wang et al. (Wang et al., 2023b) propose the Average Recall of Collaborative View (ARCV) metric, which measures recall for agents invisible from a single-vehicle perspective but detectable through collaboration. Our evaluation method follows this approach by categorizing objects into three groups: fully visible, partially visible, and fully invisible. We assess the perception performance of algorithms across these categories, first without collaboration and then with V2X collaboration, to measure the improvement provided by CP systems.

Apart from perception metrics, communication cost is also critical. In our evaluation method, we use average message size as an effective metric to assess the communication demands of collaborative perception methods (Marez et al., 2022).

Evaluating P&P introduces additional challenges, especially in comparing traditional modular method, where detection, tracking, and prediction are conducted sequentially, and end-to-end methods, which directly process sensor data to generate perception and prediction results in a unified framework. Both approaches should receive the same detection and tracking inputs for the forecasting module. For instance, Xu et al. (Xu et al., 2023) introduce a method to evaluate both traditional and end-to-end forecasting models, using the metrics summarized in Table 1. A primary metric in their approach is Mean Forecasting Average Precision (mAP_f), inspired by detection AP (Peri et al., 2022). In our work, mAP_f and mAP_{det} are the principal metrics used to assess detection and forecasting performance across different object groups: fully visible, partially visible, and fully invisible.

3.2 Evaluation in Simulation

Co-P&P is a complex multi-agent system influenced by factors such as localization error and communication constraints. To evaluate the robustness of this approach, ablation studies on key factors are essential. Simulation provides a practical solution, as it offers a fully controlled environment for testing. In our work, we use simulation to conduct various ablation studies to assess Co-P&P's performance under different conditions, including localization error, latency, and the number of CAVs. This process ensures the scalability of our approach across diverse real-world scenarios. Future research will benefit from advanced co-simulators that support realistic communication and sensor data for even more comprehensive testing.

3.3 Evaluation with Real-World Dataset

Benchmarking perception algorithms on real-world datasets is a standard approach for evaluating and comparing methods, as real-world data offers a higher degree of realism. For Co-P&P research, DAIR-V2X-Seq (Yu et al., 2023) is a useful dataset, containing 7,500 cooperative frames with infrastructure and vehicle-side images and point clouds. However, P&P relies heavily on machine learning, which requires large-scale dataset. The scale of DAIR-V2X-Seq remains limited for training larger ML models. To advance Co-P&P research, creating a more extensive CP dataset is crucial, and it is one of our primary goals for future work.

4 CHALLENGES

While Co-P&P has significant potential to enhance vehicle awareness in dynamic traffic environments by reducing accumulated errors and addressing occlusion issues, its real-world implementation faces several challenges. This section outlines key challenges in deploying Co-P&P.

- **Localization Errors:** Effective sensor data fusion requires aligning all data in a shared coordinate system, which depends on precise vehicle localization. However, GNSS-based localization typically varies in accuracy from 1 to 3 meters, leading to potential misalignments that can significantly impair data fusion. Addressing these pose errors is essential for accurate collaborative scene completion in our framework.
- **Asynchronous:** Collaborative scene completion becomes more complex due to asynchronous observations from multiple agents. To accurately reconstruct a current scene frame, input from other perspectives is often necessary. However, these inputs are frequently asynchronous with the ego vehicle's observations, causing inconsistencies in the positions of dynamic objects. Developing methods to handle asynchronous data effectively is critical for accurate scene completion.
- **Domain Shift:** In real-world traffic, vehicles from various manufacturers may be equipped with different types of LiDAR sensors, such as rotating and oscillating LiDARs. Variations in scan patterns lead to distinct data representations across sensors, introducing domain shifts that can disrupt the perception pipeline (Xiang et al., 2023; Liu et al., 2024). To prevent performance degradation, it is crucial to develop methods for completing the LiDAR scene using each sensor's unique data representation.
- **Dependency on Large-Scale Labeled Dataset:** The P&P module employs a unified neural network without hand-crafted processing steps, such as Non-Maximum Suppression (NMS). This high degree of neural network reliance increases data demands during model training. Similar to end-to-end driving models, end-to-end P&P models require large datasets. Reducing dependency on annotated data is essential to streamline P&P deployment, presenting a critical area for further investigation.

5 CONCLUSION

In this paper, we introduced a conceptual framework for Co-P&P, which comprises collaborative scene completion and P&P module. By decoupling V2X collaboration from perception, the framework enables separate training and validation of the two modules, supporting scalable deployment in real-world settings. A significant challenge in collaborative scene completion is bridging the domain gap between different LiDAR sensors, which we propose to address using a unified intermediate representation format, similar to that used in 3D reconstruction. After revisiting evaluation methods in CP, we emphasize that evaluating CP performance on objects at different visibility level provides valuable insights, particularly for objects that are fully invisible from ego view but visible from collaborative perspective. This metric highlights CP's potential to address visual occlusion, which should be considered a primary motivation for CP. Additionally, we discuss the challenges and open questions surrounding Co-P&P.

This conceptual framework serves as a high-level architecture for Co-P&P, with detailed implementation of each component to follow in future work. In addition to developing novel modules for collaborative scene completion and P&P, creating a large-scale dataset is essential to advance this field. We plan to develop a large-scale dataset supporting a range of CP tasks, including detection, tracking, and motion prediction.

REFERENCES

Abboud, K., Omar, H. A., and Zhuang, W. (2016). Interworking of dsrc and cellular network technologies for V2X communications: A survey. *IEEE Transactions on Vehicular Technology*, 65(12):9457–9470.

Bharilya, V. and Kumar, N. (2023). Machine learning for autonomous vehicle's trajectory prediction: A comprehensive survey, challenges, and future research directions. *Vehicular Communications*, 46:100733.

Chalvatzaras, A., Pratikakis, I., and Amanatiadis, A. A. (2022). A survey on map-based localization techniques for autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 8(2):1574–1596.

Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., and Fu, S. (2019). F-Cooper: Feature-based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing (SEC)*, pages 88–100. ACM.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al. (2021). Large-scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719. IEEE.

Gu, J., Hu, C., Zhang, T., Chen, X., Wang, Y., Wang, Y., and Zhao, H. (2023). ViP3D: End-to-end visual trajectory prediction via 3D agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5496–5506. IEEE.

Han, Y., Zhang, H., Li, H., Jin, Y., Lang, C., and Li, Y. (2023). Collaborative perception in autonomous driv-

ing: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine*.

Harding, J., Powell, G., Yoon, R., Fikentscher, J., Doyle, C., Sade, D., Lukuc, M., Simons, J., and Wang, J. (2014). Vehicle-to-vehicle communications: Readiness of V2V technology for application. Technical Report DOT HS 812 014, United States National Highway Traffic Safety Administration.

Hu, Y., Lu, Y., Xu, R., Xie, W., Chen, S., and Wang, Y. (2023). Collaboration helps camera overtake lidar in 3D detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252. IEEE.

Huang, T., Liu, J., Zhou, X., Nguyen, D. C., Azghadi, M. R., Xia, Y., Han, Q.-L., and Sun, S. (2023). V2X cooperative perception for autonomous driving: Recent advances and challenges. *arXiv preprint arXiv:2310.03525*.

Keen, H. E. and Berns, K. (2020). Generation of elevation maps for planning and navigation of vehicles in rough natural terrain. In Berns, K. and Görges, D., editors, *Advances in Service and Industrial Robotics*, pages 488–495, Cham. Springer International Publishing.

Keen, H. E. and Berns, K. (2023). Probabilistic fusion of surface and underwater maps in a shallow water environment. In Petrič, T., Ude, A., and Žlajpah, L., editors, *Advances in Service and Industrial Robotics*, pages 195–202, Cham. Springer Nature Switzerland.

Kenney, J. B. (2011). Dedicated short-range communications (DSRC) standards in the united states. *Proceedings of the IEEE*, 99(7):1162–1182.

Li, Y., Zhang, J., Ma, D., Wang, Y., and Feng, C. (2022). Multi-robot scene completion: Towards task-agnostic collaborative perception. In *Conference on Robot Learning (CoRL)*. PMLR.

Liang, M., Yang, B., Zeng, W., Chen, Y., Hu, R., Casas, S., and Urtasun, R. (2020). PnPNet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11553–11562. IEEE.

Liu, H., Gu, Z., Wang, C., Wang, P., and Vukobratovic, D. (2023). A lidar semantic segmentation framework for the cooperative vehicle-infrastructure system. In *Proceedings of the 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*, pages 1–5. IEEE.

Liu, Y., Sun, B., Li, Y., Hu, Y., and Wang, F.-Y. (2024). HPL-ViT: A unified perception framework for heterogeneous parallel lidars in V2V. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

Marez, D., Nans, L., and Borden, S. (2022). Bandwidth constrained cooperative object detection in images. In *Artificial Intelligence and Machine Learning in Defense Applications IV*, volume 12276, pages 128–140. SPIE.

Ochieng, W. and Sauer, K. (2002). Urban road transport navigation: Performance of the global positioning system after selective availability. *Transportation Research Part C: Emerging Technologies*, 10(3):171–187.

Peri, N., Luiten, J., Li, M., Osep, A., Leal-Taixé, L., and Ramanan, D. (2022). Forecasting from lidar via future object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17202–17211. IEEE.

Triess, L. T., Dreissig, M., Rist, C. B., and Zöllner, J. M. (2021). A survey on deep domain adaptation for lidar perception. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 350–357. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, B., Zhang, L., Wang, Z., Zhao, Y., and Zhou, T. (2023a). CoRe: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8676–8686. IEEE.

Wang, T., Chen, G., Chen, K., Liu, Z., Zhang, B., Knoll, A., and Jiang, C. (2023b). UMC: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8153–8162. IEEE.

Wang, T., Kim, S., Jiang, W., Xie, E., Ge, C., Chen, J., Li, Z., and Luo, P. (2024). DeepAccident: A motion and accident prediction benchmark for V2X autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5599–5606.

Wang, T.-H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., and Urtasun, R. (2020). V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12347, pages 605–621. Springer International Publishing.

Xiang, L., Yin, J., Li, W., Xu, C.-Z., Yang, R., and Shen, J. (2023). DI-V2X: Learning domain-invariant representation for vehicle-infrastructure collaborative 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

Xu, Y., Chambon, L., Zablocki, É., Chen, M., Alahi, A., Cord, M., and Pérez, P. (2023). Towards motion forecasting with real-world perception inputs: Are end-to-end approaches competitive? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18428–18435. IEEE.

Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., Song, J., Yuan, J., Luo, P., and Nie, Z. (2023). V2X-Seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5486–5495. IEEE.

Zhou, Y. and Tuzel, O. (2018). VoxelNet: End-to-end learning for point cloud based 3D object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499. IEEE.