Improving Vulnerable Road-Users Detection through Hybrid Collaborative Perception and Detection Refinement

Selma Oubouabdellah, Minh-Quan Dao, Ezio Malis, Elwan Héry, Julien Moreau, Vincent Frémont

Improving Vulnerable Road-Users Detection through Hybrid Collaborative Perception and Detection Refinement

Selma Oubouabdellah[1,3], Minh-Quan Dao[2], Ezio Malis[2], Elwan Héry[1], Julien Moreau[3], Vincent Frémont[1]
[1] École Centrale de Nantes, [2] INRIA, [3] Université de Technologie de Compiègne
[1]first-name.lastname@ec-nantes.fr, [2]first-name.lastname@inria.fr, [3]first-name.lastname@hds.utc.fr

***Abstract*— Ensuring the safety of autonomous vehicles in complex urban environments critically depends on accurate 3D object detection. While LiDAR sensors provide reliable depth information, their effectiveness is limited by sparsity at long distances and occlusions, particularly in intersection scenarios. Collaborative perception addresses these challenges by enabling information sharing among vehicles and infrastructure sensors, with intermediate fusion offering a balance between communication efficiency and detection accuracy. However, existing collaborative perception frameworks exhibit a notable performance gap between detecting vehicles and vulnerable road users such as cyclists and pedestrians. In this work, we propose a novel hybrid collaboration framework designed to reduce this gap. Our method leverages late-stage information from communicating agents to augment the ego agent's point cloud, then applies a standard intermediate fusion strategy, followed by a refinement stage that further improves the detection accuracy of various objects. Experiments on the Mixed Signals dataset demonstrate that our approach sets a new state-of-the-art in the detection of vulnerable road users in urban V2X scenarios.**

Index Terms— Hybrid collaborative perception, V2X communication, 3D object detection, vulnerable road users, refinement stage, urban autonomous driving.

I. Introduction

Autonomous navigation systems offer a promising solution for managing and optimizing transportation in densely populated areas. However, ensuring the safety of such systems critically depends on understanding the vehicle's environment. The most important task to achieve this understanding in urban areas is 3D object detection [1], which enables autonomous vehicles to predict the locations, sizes and categories of surrounding 3D objects, thereby preventing collisions.

Among the various sensors used in autonomous driving, LiDAR stands out as the most prevalent due to its accurate metric information and superior depth estimation capabilities. However, its effectiveness decreases with distance, as the resulting point clouds become sparser, limiting the detection of small or distant objects such as pedestrians. Additionally, its on-vehicle placement prevents it from seeing through occlusions, posing challenges in complex urban environments, particularly at intersections.

Collaborative perception helps overcome the limitations of LiDAR by sharing information between vehicles and infrastructure sensors [2]. Each agent provides a unique perspective, especially roadside units (RSUs) which are positioned at elevated locations, thereby improving visibility in occluded areas. The method of sharing information defines the type of collaboration: Early Fusion exchanges raw sensor data, offering high accuracy but requiring significant bandwidth. Late Fusion shares processed results, making it more suitable for real-time applications, although it may reduce accuracy. Intermediate Fusion provides a balance between accuracy and communication efficiency [2], [3]. Lastly, Hybrid Fusion intelligently combines elements of the other methods. For example, it may involve sharing processed data, as in Late Fusion, but integrating it earlier in the processing pipeline to improve performance [4], [5].

State-of-the-art methods typically demonstrate varying levels of detection accuracy, measured by mean average precision (mAP), across different object classes. In particular, a noticeable performance gap exists between the detection of vehicles like cars, and road users such as cyclists and pedestrians. For example, with V2X-ViT [3] we obtain approximately a 10% higher mAP for cars than for bicycles and pedestrians [6], increasing the risk to vulnerable road users whose behavior is inherently unpredictable and more challenging to handle.

To enhance the safety of autonomous driving in crowded urban areas, it is essential to address the performance gap issue. This work focuses on improving the detection accuracy of vulnerable road users, specifically bicycles and pedestrians, using the Mixed Signals dataset [6], which provides real-world Vehicle-to-Everything (V2X) data from urban environments with strong representation of the targeted classes.

The main contributions are:

- A novel hybrid framework that first augments the ego vehicle's point cloud with late-stage data from communicating agents, following the approach in [5], and then performs standard intermediate fusion as in [2], [3], producing an initial set of 3D detection proposals.
- A second-stage refinement module that improves the initial proposals by pooling Region of Interest (RoI) features from the collaborative Bird's Eye View (BEV) feature map and integrating complementary late-stage

data from neighboring agents.

Experiments on the Mixed Signals dataset show that our method achieves state-of-the-art detection of vulnerable road users, improving Bike and Pedestrian Average Precision (AP) by 12% over Attentive Fusion at a 0.7 Intersection over Union (IoU) threshold, and by 6% and 4%, respectively, when integrated with V2X-ViT. At a 0.5 IoU threshold, Bike and Pedestrian AP exceed 90% and 80% of Vehicle AP, respectively.

II. Related Works

A. Collaborative Perception

To select the most appropriate collaboration strategy for our application, it is crucial to consider key factors such as detection accuracy and bandwidth efficiency, as they directly impact both safety and real-time responsiveness. Given the high bandwidth demands of early fusion and the substantial information loss associated with late fusion, our approach will focus on intermediate and hybrid fusion strategies.

The general scheme of intermediate fusion can be divided into: metadata sharing and feature extraction, feature compression and sharing, feature fusion, and prediction header. Most methods differ primarily in their feature fusion strategies. While some approaches have achieved strong performance using graph neural networks [7], [8], others have incorporated attention mechanisms to better capture feature relationships [9]. AttFusion [2] introduced a single-head self-attention fusion module to model interactions within specific areas of the feature maps. V2X-ViT [3] proposed a heterogeneous multi-agent attention module designed to learn the distinct relationships between vehicle and infrastructure agents, and a multi-scale window attention module to capture long-range spatial interaction on high-resolution detection.

These fusion methods have achieved state-of-the-art performance while remaining suitable for real-time applications. However, results consistently reveal a performance gap between the car class and other road users classes [6]. To address this issue, our work integrates the hybrid fusion approach LaLy [5] into these intermediate fusion frameworks. LaLy utilizes late data, 3D bounding boxes, from other agents to augment the ego agent's point cloud in regions of interest (RoI). This augmentation provides additional context, aiming to improve the perception accuracy for various object classes.

B. Refinement Stage

Several studies have investigated the integration of a refinement stage to 3D object detection, demonstrating its effectiveness in enhancing bounding box localization and classification, particularly for small or distant objects [10]. PointRCNN [11] segments foreground points and generates 3D proposals, which are subsequently refined using semantic features. Building on this, Part-A^2 [12] introduces part-aware supervision to enhance intra-object feature learning. PV-RCNN [13] combines voxel-based and point-based features, leveraging RoI-grid pooling for proposal refinement. Voxel R-CNN [14] further advances voxel-based detection by proposing Voxel RoI Pooling, enabling efficient extraction of refined features directly from sparse voxel representations, effectively reducing unnecessary computation.

To improve the detection precision of cyclists and pedestrians, which occupy only a few pixels in the BEV feature map, we introduce a refinement stage after the hybrid fusion stage, inspired by Voxel R-CNN [14]. While Voxel R-CNN enhances BEV features by pooling from multi-resolution voxel grids generated at different backbone stages, we argue that the loss of fine spatial information is compensated through feature fusion among multiple vehicles and RSUs. We therefore refine detections directly on the fused BEV feature map, eliminating the need to construct additional voxel grids. This choice enables a more efficient pooling strategy that queries nearest neighbors in the 2D plane, avoiding the higher computational cost of 3D neighbor searches required by Voxel R-CNN.

III. Method

A. Overview

Fig. 1 provides an overview of the method proposed in this work. It consists of three main components, which are described in the following sections. Section III.B details the intermediate collaboration stage, where initial detections are produced through the collaborative fusion of feature maps shared by communicating agents. The output of this stage is then passed to the refinement stage, described in Section III.C, which further improves the detection results by refining the centers and sizes of the predicted bounding boxes. Finally, Section III.D presents the integration of Laly Fusion into both the intermediate collaboration and refinement stages, enabling more accurate detection by effectively adding context through the fusion of multi-source information at each stage.

B. First Stage - Intermediate Collaboration

Before intermediate fusion can occur, the raw point cloud must first be processed and communicated between agents. To this end, the 3D space is discretized into structured representations, enabling feature extraction and organization into a two-dimensional pseudo-image. A multi-stage BEV backbone with progressive downsampling and upsampling further refines the spatial feature representation to produce the BEV feature map that is shared among agents.

Our general approach shown in Fig. 1 is designed to be modular with respect to the intermediate fusion strategy. The framework remains agnostic to the specific feature fusion method, enabling the incorporation of any technique capable of aggregating feature maps from multiple agents. To validate the generality and effectiveness of our method, we implement and evaluate both Attentive Fusion [2] and V2X-ViT [3] as intermediate fusion modules.

Attentive Fusion models collaborative perception as a spatial graph, where nodes correspond to vehicles and edges represent communication links. Feature fusion is achieved by applying self-attention across spatially aligned features, using a local graph at each spatial location to capture interactions

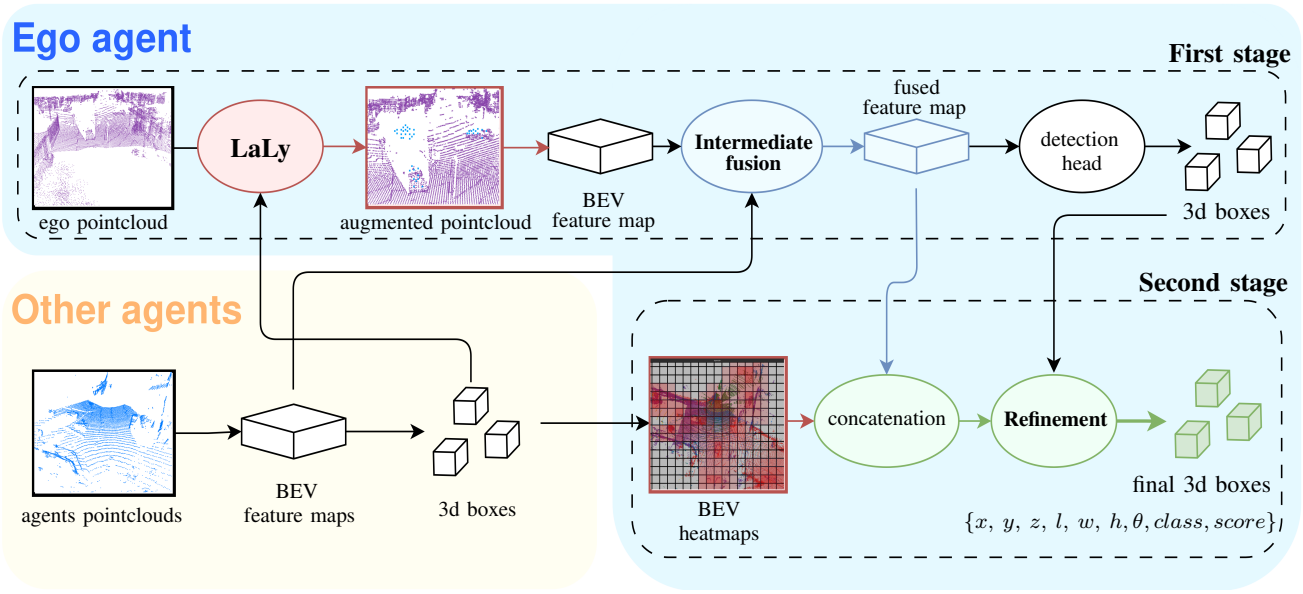

Fig. 1: Overview of the proposed method. The first stage performs hybrid collaborative fusion, where the ego agent's point cloud is augmented with detections from other agents before feature extraction and intermediate fusion. The second stage refines the initial ego detections using the collaborative BEV feature map and the detection heatmaps from other agents.

between connected vehicles. This produces a collaborative feature map that integrates complementary observations.

V2X-ViT models cooperative perception as a heterogeneous graph, where nodes represent vehicles or RSUs, and edges encode different interaction types. It introduces a heterogeneous multi-agent self-attention module that captures inter-agent relationships by considering node and edge types, while restricting attention to spatially aligned features across agents to preserve spatial structure. To improve robustness against localization errors, a multi-scale window attention module aggregates spatial features across multiple window sizes and fuses them adaptively. By stacking multiple blocks of these modules, V2X-ViT progressively refines spatial and relational feature representations while maintaining high-resolution inputs, producing a robust collaborative feature map for object detection.

The resulting feature map is decoded by an anchor-based detection head, which predicts object classes and bounding box parameters. Standard post-processing operations, including non-maximum suppression, are then applied to obtain the final set of initial detections, which serve as input to the subsequent refinement stage.

C. Second Stage - Refinement

After obtaining the collaborative feature map and initial 3D bounding boxes, we refine the detections as illustrated in Fig. 2.

The refinement stage takes as input a dense BEV feature map of size (C, H, W), where C is the number of channels and H, W are the height and width. For each initial bounding box proposal $\mathcal{B}$ from the first stage, we compute a feature vector as follows. We first sample a grid of $N \times M$ points from the top face of each proposal. Each grid point $\mathbf{p}_i$ is

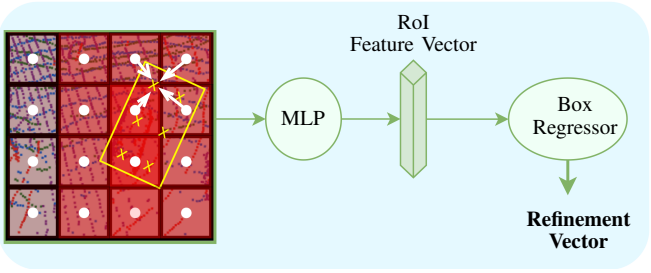

Fig. 2: Refinement Pipeline: cropped BEV features are processed through an MLP, aggregated into a feature vector, and used for accurate bounding box refinement.

projected to the BEV feature map coordinates using:

$$\text{pixel index} = \frac{\mathbf{x} - \mathbf{x}_{\text{min}}}{\text{stride} \times \text{voxel size}} \tag{1}$$

where $\mathbf{x}$ is the 3D coordinate of $\mathbf{p}_i$ and $\mathbf{x}_{\text{min}}$ is the minimum BEV map range. The feature $\mathbf{f}_i \in \mathbb{R}^C$ for each grid point is obtained by bilinear interpolation from the BEV feature map, then transformed to a higher dimension C' using a Multi-Layer Perceptron (MLP). The features from all grid points are concatenated to form a matrix $\mathbf{f} \in \mathbb{R}^{C' \times (N \cdot M)}$. Max pooling along the second dimension ensures permutation invariance, resulting in the feature vector $\mathbf{f}_\mathcal{B}$ for proposal $\mathcal{B}$. Two MLPs decode $\mathbf{f}_\mathcal{B}$ into a new confidence score s and a refinement vector $\mathbf{\Delta}$.

During training, we sample a fixed number of proposals $\{\mathcal{B}_k | k = 1, ..., K\}$ from the first stage. To generate targets for the confidence score s, proposals are matched to ground truth if their BEV IoU exceeds a threshold. This matching is non-exclusive, allowing a ground truth to match multiple

proposals. The target for s is 1 for matched proposals and 0 otherwise. For each matched proposal $\mathcal{B}_m$, the ground truth $\mathcal{G}_m$ with the highest IoU is used to compute the refinement target $\mathbf{\Delta} = [\delta_x, \delta_y, \delta_z, \delta_h, \delta_w, \delta_l, \delta_\theta]$. The center offset is normalized by the diagonal of $\mathcal{B}_m$:

$$\delta_x = \frac{x_G - x_B}{\sqrt{w_B^2 + l_B^2}} \tag{2}$$

where x_G and x_B are the x-coordinates of $\mathcal{G}_m$ and $\mathcal{B}_m$, and w_B, l_B are the width and length of $\mathcal{B}_m$. The dimension refinement is defined as the log ratio:

$$\delta_l = \log \frac{l_G}{l_B} \tag{3}$$

For unmatched proposals, the refinement target $\mathbf{\Delta}$ is set to null and their loss contribution is zero.

During inference, proposals that have sufficiently high confidence score predicted by the first stage are kept for refinement.

D. Integration of Laly Fusion

We adopt Laly Fusion in a broader sense than [5], fusing objects detected by other agents (vehicles or RSU) with information from the ego vehicle's LiDAR and perception model.

Integrating Laly Fusion into the first stage: To incorporate objects detected by other agents, we follow the original Laly Fusion framework [5]. Detected 3D bounding boxes are first transformed into the ego vehicle's coordinate system. Each bounding box is treated as an augmented point, with coordinates set to the box center and features including dimensions, heading, confidence score, and class. These augmented points are combined with the ego vehicle's point cloud, after padding the ego points' features with zeros to match dimensions. The aggregated point cloud is then input to the first stage.

Integrating Laly Fusion into the refinement stage: For the refinement stage, each detected object is converted into a 2D Gaussian heat map $\mathbf{H} \in \mathbb{R}^{H \times W}$ using the CenterNet formulation [15]:

$$\mathbf{H}[x, y] = \exp\left(-\frac{(x - \tilde{c}_x)^2 + (y - \tilde{c}_y)^2}{2\sigma^2}\right) \tag{4}$$

where $[\tilde{c}_x, \tilde{c}_y]$ is the pixel coordinate of the bounding box center, and σ is determined from the box's width and length as in [16].

Heat maps for objects of the same class are aggregated by taking the maximum value at each pixel. These class-specific heat maps are then concatenated to the feature map from the first stage along the channel dimension.

IV. Experiments

A. Experiment Setup

Dataset: We evaluate our method using the Mixed Signals dataset [6], collected by three connected vehicles and an RSU positioned 2.5 meters above ground at a Sydney intersection. Each vehicle is equipped with an OS1 128-beam LiDAR, while the RSU has both an OS-Dome 128-beam and an OS1 64-beam LiDAR. All sensors are synchronized at 10 Hz.

The dataset contains 37 scenes, with 33 used for training and 4 for testing. The training and test sets include 9,553 and 1,164 samples, respectively. Each sample includes point clouds from the RSU's two LiDARs and from vehicles within 70 meters of the RSU. Annotations are provided as 3D bounding boxes for three classes: Vehicle, Bike, and Pedestrian.

Evaluation protocol: We follow the Mixed Signals dataset protocol, transforming predictions and ground truth into the RSU coordinate system and filtering out objects outside the range of $[-51.2, 51.2]$ meters along both x and y axes. Detection performance is measured by Average Precision (AP).

To compute AP, detected objects are matched to ground truth based on their intersection over union in the bird's-eye view. A detection is considered a true positive if its IoU with a ground truth exceeds a specified threshold. Unmatched predictions are counted as false positives, and unmatched ground truths as false negatives. For comparisons in Sec. IV-B, we use IoU thresholds of 0.3, 0.5, and 0.7. In this section, we also report the mean AP at these three IoU thresholds which is the AP averaged over three classes. For ablation studies in Sec. IV-C, we use only the 0.7 threshold.

Implementation details: We conduct our experiments using the OpenCOOD framework [2], which implements various collaborative methods. For fair comparison, all models use PointPillar [17] as the backbone. The training is conducted on a cluster of H100 GPUs.

To demonstrate the effectiveness and compatibility of our method with different intermediate fusion approaches, we instantiate two models using Attentive Fusion [2] and V2X-ViT [18] as the first stage, referred to as **Ours-A** and **Ours-V**, respectively. The hyperparameters for Attentive Fusion and V2X-ViT are set identical to their official implementations[1]. The first stage of Ours-A and Ours-V is initialized with weights from Attentive Fusion and V2X-ViT, trained for 20 and 30 epochs, respectively, without the refinement stage and Laly fusion. The refinement stage weights are initialized randomly. We then train Ours-A and Ours-V end-to-end for 20 epochs using Adam optimizer [19]. The initial learning rates are set to 0.0002 for Ours-A and 0.001 for Ours-V. For Ours-A, the learning rate is reduced by a factor of 10 after 7 and 12 epochs, while for Ours-V, it is reduced by a factor of 10 after 15 epochs. To train the second stage, we set the threshold for matching proposals generated by the first stage to ground truth to 0.3.

B. Quantitative Evaluation

Tab. I compares the performance of our models, Ours-A and Ours-V, with Early Fusion, Intermediate Fusion, and Laly Fusion. The Intermediate Fusion methods include F-Cooper [20], V2VNet [21], Where2comm [22], Attentive Fusion [2], and V2X-ViT [18].

[1] https://github.com/DerrickXuNu/OpenCOOD

TABLE I: Comparison of AP measured at three IoU thresholds between our method and prior works. The best and second best result in each column are highlighted by **boldface** and underline, respectively.

	Vehicle @ IoU			Bike @ IoU			Pedestrian @ IoU			mean AP @ IoU		
	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
No Fusion	0.17	0.17	0.17	0.18	0.18	0.17	0.25	0.21	0.05	0.20	0.19	0.13
Early Fusion	0.71	0.71	0.71	0.63	0.62	0.62	0.69	0.61	0.24	0.68	0.65	0.52
F-Cooper [20]	0.78	0.78	0.78	0.75	0.75	0.62	0.73	0.70	0.40	0.75	0.74	0.60
V2VNet [21]	0.72	0.72	0.72	0.70	0.69	0.69	0.42	0.32	0.13	0.61	0.58	0.51
Where2comm [22]	0.77	0.77	0.77	0.71	0.71	0.62	0.43	0.36	0.16	0.64	0.61	0.52
Attentive Fusion [2]	0.83	0.83	0.83	0.65	0.65	0.65	0.75	0.68	0.24	0.74	0.72	0.57
V2X-ViT [18]	**0.85**	**0.85**	**0.85**	0.74	0.74	0.66	0.77	0.75	0.54	0.79	0.78	0.68
Laly Fusion [5]	0.61	0.61	0.61	0.68	0.68	0.68	0.69	0.62	0.28	0.66	0.64	0.52
Ours-A	0.81	0.81	0.81	0.78	0.77	**0.77**	0.77	0.71	0.36	0.79	0.76	0.65
Ours-V	0.84	0.84	0.84	**0.79**	**0.78**	0.72	**0.81**	**0.80**	**0.58**	**0.81**	**0.81**	**0.71**

Integrating our method with V2X-ViT (Ours-V) achieves state-of-the-art results in detecting vulnerable road users, specifically the Bike and Pedestrian classes. Compared to Attentive Fusion, our method (Ours-A) improves Bike and Pedestrian detection by 12% AP at the challenging 0.7 IoU threshold. For V2X-ViT, our approach yields improvements of 6% AP for Bike and 4% AP for Pedestrian.

Despite the class imbalance in the training set and without any class-specific modules, our method achieves comparable precision for Vehicle, Bike, and Pedestrian. At a 0.5 IoU threshold, the AP for Bike and Pedestrian exceeds 90% and 80% of the Vehicle AP, respectively, for both Ours-A and Ours-V.

However, the AP for Pedestrian drops significantly as the IoU threshold increases from 0.5 to 0.7. This is due to the large pixel size in the Bird-Eye View images produced at the end of the first stage—0.4 meters for Attentive Fusion and 0.8 meters for V2X-ViT. As a result, a pedestrian often occupies less than a pixel, making precise localization at the 0.7 IoU threshold challenging.

C. Ablation Studies

To assess the contribution of key components in our method—including the refinement stage and the integration of Laly Fusion in both stages—we conduct ablation experiments starting from Attentive Fusion and progressively adding these components. The results are shown in Tab. II.

Each component differently contributes to detection performance. Laly Fusion in the first stage improves Bike AP by 14% and Pedestrian AP by 8% due to denser lidar points. Adding the refinement stage further improves Pedestrian AP by 10% through precise localization and dimension adjustments of initial proposals, though it slightly reduces Bike AP by 1%. Combining all components yields an overall AP increase of 12% for Bikes and Pedestrians, with a minor 2% drop for Vehicles, likely due to hyperparameter tuning rather than an inherent limitation of our proposed pipeline.

D. Qualitative Evaluation

Fig. 3 shows the visual comparison between Attentive Fusion and Ours-A. It can be seen that our model is able to detect a number of Bike and Pedestrian that are missed by Attentive Fusion. In orther words, our model produces less false negatives in detecting Bike and Pedestrian. This explains the higher precision achieved by our model.

TABLE II: Ablation study: relative improvement of AP measured at 0.7 IoU threshold of Attentive Fusion brought by different modules of our method

1st Stage	Refine. Stage	Laly in 1st	Laly in 2nd	Vehicle	Bike	Pedestrian
✓				0.83	0.65	0.24
✓		✓		-0.02	+0.14	+0.08
✓	✓			+0.01	+0.00	+0.05
✓	✓		✓	+0.01	-0.01	+0.10
✓	✓	✓	✓	-0.02	+0.12	+0.12

V. Conclusion

This work addresses the precision gap in V2X object detection between vehicles and vulnerable road users. We proposed a two-stage hybrid framework that first augments the ego vehicle's point cloud with late-fusion detections from neighboring agents, then performs intermediate feature fusion to generate initial 3D proposals, and finally refines these proposals through RoI pooling over the collaborative BEV feature map. Evaluated on Mixed Signals dataset, our method improves Bike and Pedestrian AP by 12% over Attentive Fusion at a 0.7 IoU threshold, and by 6% and 4%, respectively, when integrated with V2X-ViT. At a 0.5 IoU threshold, Bike and Pedestrian AP exceed 90% and 80% of Vehicle AP, respectively. These results establish a new state of the art for vulnerable road users detection in complex urban environments. In future work, we plan to incorporate camera data into the intermediate fusion stage to enhance the precision of the perception system [1]. Additionally, we aim to extend the pipeline to include semantic segmentation, providing richer scene understanding and enabling safer navigation in complex urban environments.

VI. Acknowledgement

This work has been carried out and funded in the framework of the ANNAPOLIS project managed by the Na-

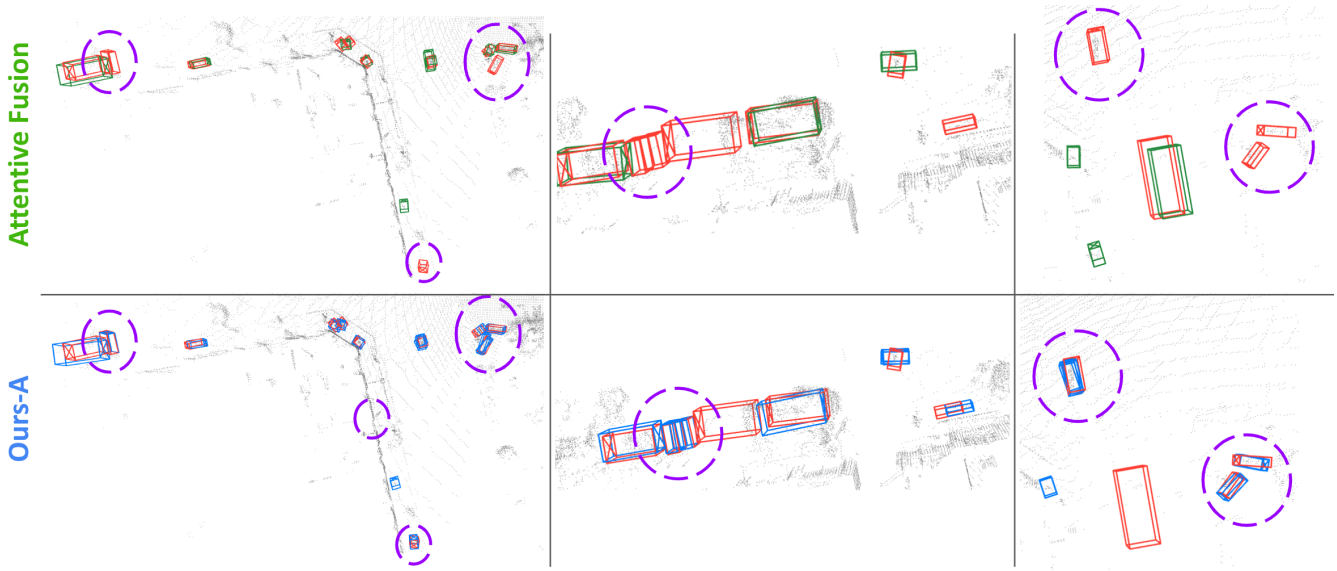

Fig. 3: Qualitative comparison between Attentive Fusion and Ours-A in three data samples of the test set of Mixed Signals. Ground truths are represented by red. Object detected by Attentive Fusion and Ours-A are respectively denoted by green and blue. Purple dashed circles highlight ground truths that are missed by Attentive Fusion but detected by Ours-A.

tional Agency for Research (ANR21-CE22-0014). As it was granted access to the HPC resources of IDRIS under the allocation 2024-AD011012128R4 made by GENCI. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

[1] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1909–1963, 2023.

[2] 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 *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.

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

[4] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1852–1864, 2020.

[5] 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 Transactions on Intelligent Transportation Systems*, 2024.

[6] K. Z. Luo, M.-Q. Dao, Z. Liu, M. Campbell, W.-L. Chao, K. Q. Weinberger, E. Malis, V. Fremont, B. Hariharan, M. Shan *et al.*, "Mixed signals: A diverse point cloud dataset for heterogeneous lidar v2x collaboration," *arXiv preprint arXiv:2502.14156*, 2025.

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

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

[9] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets and challenges," *arXiv preprint arXiv:2301.06262*, 2023.

[10] C. Shi, C. Zhang, and Y. Luo, "Structure guided proposal completion for 3d object detection," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 4462–4478.

[11] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[12] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2647–2664, 2020.

[13] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538.

[14] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.

[15] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[16] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[17] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

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

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[21] 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 *ECCV*. Springer, 2020, pp. 605–621.

[22] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022.