LFF-V2V: A Late Fusion Cooperative Framework in V2V Scenarios

Alberto Justo[1,3], Javier Araluce[1], Mario Rodriguez-Arozamena[1,2], Leonardo Gonzalez[1] and Luis Miguel Bergasa[3]

***Abstract*— Traditional perception systems in automated driving have different constraints that do not allow for complete environmental awareness. Cooperative Perception (CP) addresses these limitations by sharing information between vehicles and/or infrastructure through Vehicle-to-Everything (V2X) communications. This collaborative approach mitigates occlusions and extends sensor coverage, proving essential for Cooperative Driving Automation (CDA). However, there are remaining challenges about its application in real-world scenarios, such as CP information transmission and communication degradations. In this cooperative context, Motion Prediction (MP) proves to be crucial, since it provides a scene representation of all the agents with their positions, velocities and future trajectories. Thus, shared information between agents can improve each agent understanding of the overall scene.**

This paper introduces LFF-V2V, A Late Fusion Cooperative Framework in V2V Scenarios. It combines two late fusion methods, Non-Maximum Suppression (NMS) and Weighted Box Fusion (WBF), with a map-less Hierarchical Vector Transformer (HiVT) motion prediction model. We have conducted an extensive evaluation in two environments: CARLA simulator and the real-world V2X-Real dataset, analyzing different communication strategies. Our results demonstrate the effectiveness of CP in improving object detection and motion prediction, even in degraded environments.

I. INTRODUCTION

Over the years, perception systems have proven to be essential for automated driving applications [1], [2]. However, the single-vehicle perception paradigm, where each vehicle relies solely on its onboard sensors, limits the scope and reliability of these systems [3]. Challenges such as occlusions, limited sensor range, and computational cost, disable a complete and accurate understanding of the environment. To address these challenges, Cooperative Perception (CP) stands as one of the main solutions [4]. CP is crucial in Cooperative Driving Automation (CDA), which aims to improve the safety and flow of traffic by supporting the movement of multiple vehicles in proximity to each other [5]. Through shared information between vehicles and/or infrastructure, CP enables multi-view perspectives to create a richer, more comprehensive representation of the environment. This collaborative approach improves situational awareness, resolves occlusion issues, and extends the effective range of perception systems during CDA [6]. Despite its potential, current CP faces several challenges. One major issue is the synchronization of data across multiple agents through Vehicle-to-Everything (V2X) communications, which requires low latencies to ensure a balanced trade-off between accuracy and feasible bandwidths [7]. Moreover, data association and fusion methods [8] are a major concern for CP, as inconsistent or noisy input from different agents degrades its overall performance.

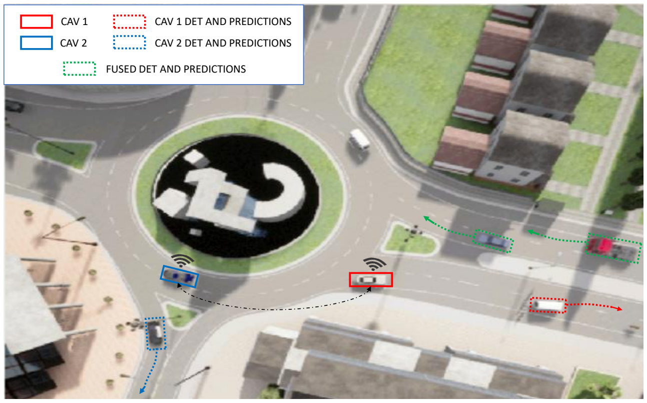

CAV 1
CAV 2
CAV 1 DET AND PREDICTIONS
CAV 2 DET AND PREDICTIONS
FUSED DET AND PREDICTIONS

Fig. 1: CARLA V2V scenario illustrating predicted trajectories based on bounding box detections. Red and blue dashed boxes and lines correspond to detections and predictions from CAV 1 and CAV 2, respectively. Detections and trajectories derived from fused detections of both CAVs are shown as green dashed boxes and lines.

Motion Prediction (MP), also known as motion forecasting, of the behavior of surrounding agents is essential for safe and efficient navigation. CP offers a powerful tool to push the limits of what is achievable with single-vehicle MP systems [9]. Thanks to V2X information exchange between agents, their object detections are enhanced. Since detections of the surrounding agents are improved, this leads to more accurate agent trajectory forecasts [10]. This improvement is particularly notable in places where occlusions can limit the effectiveness of individual sensors, like intersections or roundabouts. Thanks to these improved detections and predictions, the motion planning systems of each Connected Automated Vehicle (CAV) can make more informed and safer decisions in CDA [11].

Figure 1 illustrates a CARLA roundabout case where CP is applied, overcoming occlusions for CAV 1 and CAV 2. In this context, CAV 2 is not able to detect nor predict the vehicles that are entering the roundabout. However, thanks to V2V information exchange with CAV 1, CAV 2 can detect these vehicles and predict their trajectories. Enhanced agents detections and forecasts entering the roundabout are represented for clearer interpretation.

Our paper presents LFF-V2V, A Late Fusion Cooperative

[1] Alberto Justo, Javier Araluce, Mario Rodriguez-Arozamena and Leonardo González are with TECNALIA, Basque Research and Technology Alliance (BRTA), 48160 Derio, Bizkaia, Spain `alberto.justo, javier.araluce, mario.rodriguez, leonardo.gonzalez{@tecnalia.com}`

[2] Mario Rodriguez-Arozamena is with Department of Automatic Control and Systems Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain `mrodriguez183@ikasle.ehu.eus`

[3] Alberto Justo and Luis Miguel Bergasa are with Department of Electronics, University of Alcalá, 28805 Alcalá de Henares, Spain `alberto.justo, luism.bergasa{@uah.es}`

Framework in V2V Scenarios. This framework gathers two late fusion methods [12], [13] and HiVT-based motion prediction model [14]. We evaluated LFF-V2V in CARLA [15] and real-world dataset V2X-Real [16]. Our contributions are as follows:

- Enhanced end-to-end cooperative perception framework with map-less motion prediction model [17].
- Evaluation with two well stablished state-of-the-art late fusion techniques: Non-Maximum Supression [12] and Weighted Box Fusion [13].
- Extensive analysis of the framework in CARLA [15] and the new real-world dataset V2X-Real [16], with different communication strategies.

II. Related Works

A. Is Cooperative Perception feasible?

The feasibility of Cooperative Perception (CP) has been a subject of extensive research, particularly in the context of automated driving. The effectiveness of CP depends on several factors, including: cooperative datasets, collaboration modalities, resilience to system degradations and applicability in real-world scenarios using standardized Vehicle-to-Everything (V2X) communications [18].

To enable CP, various datasets have been developed [19]–[23], capturing multi-agent interactions under different conditions, with synchronized multi-view sensor data for benchmarking object detection and motion prediction models. However, practical CP remains not present in these datasets, since they do not count on communications for evaluation. CP can take various forms, ranging from early sensor fusion [24], intermediate fusion [25], [26] and late fusion [12], [13]. Because of communication feasibility, the last two approaches are currently the most used. Yet, given the actual ETSI standards [18] for V2X communications, Collective Perception Messages (CPM) are suited for late fusion, since they only contain post-processed bounding box detections of each agent. Thus, in most of practical CP implementations, late fusion is the method used [27]. Real-world CP implementations face degradation factors like communication delays, localization inaccuracies, and sensor noise [28]. Studies like [29] have evaluated package loss and travel time in a real-world cooperative environment. However, they do not analyze their degradation effect in CP. Since we want to analyze time-loss degradation in CP, we evaluate LFF-V2V in CARLA [15] and V2X-Real [16] in different communication contexts.

B. Motion Prediction through Cooperative Perception

MP is a critical component of automated driving, allowing vehicles to anticipate the future trajectories of surrounding agents for safe and efficient navigation [2]. Traditional MP approaches often rely on non-cooperative environments that use high-definition maps [14], graph convolution methods [30] or motion transformers [31]. Cooperative MP improves the social context between agents, as they can be more aware of the behavior of their surroundings. Recent cooperative solutions in MP, such as [32], extend the capabilities of single-agent MP by integrating multi-agent observations without pre-mapped information. Our previous work in [17] provides an enhanced MP framework through V2V collaboration. However, both studies do not count on possible degradations that can affect cooperative MP, like object detection errors or time-loss delays. Recent works like [7] highlight the importance of robust data association and timestamp alignment to mitigate these challenges. From our point of view, MP needs to be adapted and evaluated to varying levels of CP degradation without compromising performance.

III. LFF-V2V Framework

The proposed LFF-V2V framework introduces a modular pipeline for CP in V2V scenarios. Our framework aims to enhance object detection and trajectory forecasting in different collaborative driving environments. Figure 2 illustrates the architecture of LFF-V2V, showing the integration of each submodule and their interaction within the overall system.

A. Object Detection and Tracking

In our framework, both CAVs share the same object detection and multi-object tracking core. This module of our LFF-V2V framework is only included in CARLA. Detection is built on LiDAR-based PointPillars model [33] for each CAV. PointPillars model was trained on SHIFT dataset [34], which provides synthetic LiDAR data suitable for automated driving applications. The output of PointPillars is then fed into the Simple Online and Realtime Tracking (SORT [35]) algorithm, which performs a tracking-by-detection approach. SORT associates detections across consecutive frames using a combination of well-established techniques, such as the Kalman Filter and the Hungarian algorithm. This combination of detection and tracking provides the necessary input data for the V2X module.

B. V2X Module

The V2X module enables the exchange of tracking data between CAVs using standardized communication protocols defined by [18]. This ensures interoperability and robust message handling in V2V scenarios through Cooperative Awareness Messages (CAM) and Collective Perception Messages (CPM). They are transmitted in formats compatible with ROS [36], carrying information about detected objects, including their positions, dimensions, orientations, and confidence scores.

To emulate real-world conditions, the V2X module incorporates synthetic noise in form of communication delays and packet loss rates, following [32]. Details about this module will be explained in the next section of our paper. It allows our framework to evaluate its robustness under degraded conditions where perception messages are not time-synchronized between agents.

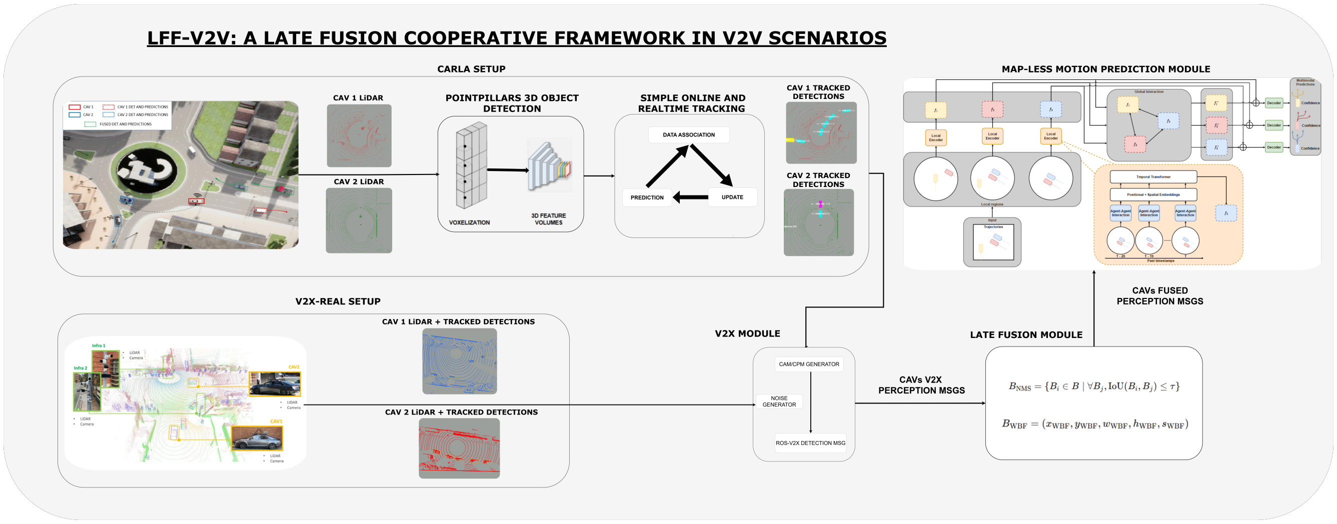

LFF-V2V: A LATE FUSION COOPERATIVE FRAMEWORK IN V2V SCENARIOS
CARLA SETUP
CAV 1
CAV 2
CAV 1 DET AND PREDICTIONS
CAV 2 DET AND PREDICTIONS
FUSED DET AND PREDICTIONS
CAV 1 LiDAR
CAV 2 LiDAR
POINTPILLARS 3D OBJECT DETECTION
VOXELIZATION
3D FEATURE VOLUMES
SIMPLE ONLINE AND REALTIME TRACKING
DATA ASSOCIATION
PREDICTION
UPDATE
CAV 1 TRACKED DETECTIONS
CAV 2 TRACKED DETECTIONS
MAP-LESS MOTION PREDICTION MODULE
Global Interaction
Local Encoder
Decoder
Confidence
Local regions
Input
Trajectories
Temporal Transformer
Positional + Spatial Embeddings
Agent-Agent Interaction
Past timestamps
CAVs FUSED PERCEPTION MSGS
V2X-REAL SETUP
Infra 1
Infra 2
LiDAR
Camera
CAV1
CAV2
CAV 1 LiDAR + TRACKED DETECTIONS
CAV 2 LiDAR + TRACKED DETECTIONS
V2X MODULE
CAM/CPM GENERATOR
NOISE GENERATOR
ROS-V2X DETECTION MSG
CAVs V2X PERCEPTION MSGS
LATE FUSION MODULE
$B_{\text{NMS}} = \{B_i \in B \mid \forall B_j, \text{IoU}(B_i, B_j) \leq \tau\}$
$B_{\text{WBF}} = (x_{\text{WBF}}, y_{\text{WBF}}, w_{\text{WBF}}, h_{\text{WBF}}, s_{\text{WBF}})$

Fig. 2: LFF-V2V: A Late Fusion Cooperative Framework in V2V Scenarios

C. Late Fusion

LFF-V2V integrates two state-of-the-art late fusion techniques to merge the detection outputs from multiple agents, which will be evaluated under time-loss degradation later in this paper.

1) Non-Maximum Suppression (NMS): In the NMS method [12], the bounding boxes are considered to belong to the same object if their Intersection-over-Union (IoU) area is greater than a predefined threshold value Th. The IoU measures the overlap between two bounding boxes in relation to their combined area, as shown in Equation 1. Once IoU is calculated for all pairs of bounding boxes, NMS discards any overlapping boxes with an IoU smaller than Th. This process suppresses redundant detections and retains only the most accurate bounding boxes. The selected boxes are represented as in Equation 2.

$$IoU(B_i, B_j) = \frac{B_i \cap B_j}{B_i \cup B_j} \tag{1}$$

$$B_{\text{NMS}} = \{B_i \in B \mid \forall B_j, \text{IoU}(B_i, B_j) \geq Th\} \tag{2}$$

Here, B_i and B_j represent the considered bounding boxes, $B_i \cap B_j$ is the intersection area, and $B_i \cup B_j$ is the union area. Th controls the trade-off between suppressing overlapping boxes and retaining valid detections.

2) Weighted Box Fusion (WBF): WBF [13] sorts bounding boxes by confidence and groups them into clusters based on an IoU threshold. For each group, the confidence score of the fused box S_{avg} is calculated as in Equation 3.

$$S_{avg} = \frac{\sum_{i=1}^{N} S_i}{N} \tag{3}$$

Here, S_i is the confidence score of each box within the cluster, and N is the total number of boxes in the cluster. The coordinates of the fused box, in Equation 4, are calculated as a weighted sum of the coordinates of the individual boxes, where the weights are their confidence scores. Same procedure applies to their width and height as shown in Equation 5.

$$X_{avg} = \frac{\sum_{i=1}^{N} S_i \cdot X_i}{\sum_{i=1}^{N} S_i}, \quad Y_{avg} = \frac{\sum_{i=1}^{N} S_i \cdot Y_i}{\sum_{i=1}^{N} S_i} \tag{4}$$

$$W_{avg} = \frac{\sum_{i=1}^{N} S_i \cdot W_i}{\sum_{i=1}^{N} S_i}, \quad H_{avg} = \frac{\sum_{i=1}^{N} S_i \cdot H_i}{\sum_{i=1}^{N} S_i} \tag{5}$$

Therefore, each fused bounding box by WBF can be represented as in Equation 6.

$$B_{\text{WBF}} = \{W_{\text{avg}}, H_{\text{avg}}, X_{\text{avg}}, Y_{\text{avg}}, S_{\text{avg}} \mid \text{IoU}(B_i, B_j) \geq Th\} \tag{6}$$

D. Map-less Motion Prediction

For MP, LFF-V2V incorporates the map-less version of the Hierarchical Vector Transformer (HiVT) model [14]. HiVT model was trained on Argoverse 1 [37], without HD-map information input as in our previous work [17]. This map-less model is optimal for V2V MP as implemented in our LFF-V2V framework. We train the model to learn the spatio-temporal relationships between the scene agents. This approach eliminates the dependency on static map information, making the system more scalable and adaptable to diverse environments without requiring specific map configurations.

Mapless inputs for this model include features such as agent positions (X,Y) and their IDs derived from detection and tracking outputs in CARLA, and ground-truth detections in V2X-Real. HiVT incorporates a local encoder that captures agent-agent interactions, ensuring a rotation-invariant representation for each agent. A global module then encodes long-range dependencies between agent-centric local representations. Finally, a multimodal future decoder predicts the trajectories of all agents in a single pass. The output is eventually represented in the local coordinate frame of the ego-vehicle.

For temporal information, we utilized 50 frames captured at a frequency of 10 Hz. This setup aligns with the configuration used in Argoverse 1, where 2 seconds of past data are considered, and the prediction horizon spans 3 seconds. In LFF-V2V, we count on the fused detections of both CAVs to enhance MP. We aim to evaluate how fused detections and communication degradation affect MP in our cooperative framework.

IV. EXPERIMENTAL SETUPS

This section describes the used experimental setups to evaluate the proposed LFF-V2V framework. We carry out our experiments in simulated and real-world environments, using CARLA and the V2X-Real dataset, respectively. Each setup is designed to assess the framework's performance under varying conditions. Both setups share the same communication module in our pipeline, with two different modes: ROS and V2X. In the ROS-based setup, the experiments are conducted asynchronously, ensuring that the entire pipeline, including detection, tracking, fusion, and motion prediction, operated at a stable frequency of 10 Hz across all submodules. This consistent frequency allows for reliable processing and evaluation of the fusion methods without introducing delays or synchronization issues. The V2X communication module simulates real-world transmission conditions by introducing a delay of 100 ms in message generation and a packet loss rate of 0.1 as in [32]. This means that 1 out of 10 CPMs are lost during transmission. The computational backbone of our experiments is: NVIDIA L40 48 GB x 2 AMD EPYC 9124 16-Core Processors.

A. CARLA

We recorded a sample V2V dataset comprising 5k frames with an average of 25 cars per frame. This data was sourced from the Town03 environment of CARLA. In our sample dataset, both CAVs are equipped with simulated 32-beams LiDAR as input, since we want our framework to operate smoothly in CARLA at 10 Hz without frame rate issues. This is the input for our PointPillars detector. Later, we apply the SORT algorithm to PointPillars' detected objects. Our CARLA experiments provide a simulation context for LFF-V2V evaluation in object detection and MP.

B. V2X-Real-V2V

V2X-Real dataset [16] provides a large-scale, real-world benchmark for cooperative perception. It was collected using a combination of 2 CAVs and 2 smart infrastructure systems. Our experimental setup uses the V2X-Real-V2V subset, focusing exclusively on V2V collaboration. It includes 17k LiDAR frames, 140k camera images, and 719k bounding boxes. However, some of its samples are not fully complete, with either subsamples of one of the CAVs missing or non-complete LiDAR recordings of the scene. For this reason, we only have been able to use 5k LiDAR frames of this subset for both CAVs, with their respective bounding boxes. For object detection and tracking, we take ground-truth detections from the dataset, which include agent IDs for data association. Our setup focuses on evaluating MP module after late fusion, without additional noise or errors from detection models. For this setup, we only use NMS as the late fusion technique, since ground-truth detections do not provide scores for using WBF. This V2X-Real-V2V integration provides a context for the evaluation of LFF-V2V MP under controlled conditions.

V. EVALUATION

This section shows the final results after conducting our experiments. For detection, we evaluated two different late fusion methods, NMS and WBF, in CARLA, with PointPillars as the base detector. In V2X-Real-V2V, object detection is not evaluated, since we take ground-truth detections directly from the dataset to assess MP afterwards. For MP, we evaluated these same fusion methods in CARLA, and only NMS for V2X-Real-V2V. Both assessments were done within different communication modes: ROS and V2X. Our evaluations show how perception performance is affected indeed in degraded cooperative environments. They include common metrics for object detection [1] and motion prediction [37]: Average Precision (AP), Minimum Average Displacement Error (minADE), Minimum Final Displacement Error (minFDE), Brier Score for minADE (b-minADE) and Brier Score for minFDE (b-minFDE). In this context, we specify object detection and MP metrics only for ”Car” class.

A. CARLA

We evaluate LFF-V2V object detection performance using two different thresholds of IoU: 0.5 and 0.7. To assess late fusion in object detection, we take ground-truth bounding boxes from CARLA agents, applying GTPC method as in [38] for each CAV: boxes filter through ray-tracing, taking as valid those which have more than 10 LiDAR points contained in them. To make a more equitable comparison between MP metrics, we decide to take the highest IoU threshold (0.7) for the late fusion methods. Thus, overall detections are more strict, hence MP errors are higher. In MP metrics, we focus on the reliability of the brier score metrics (b-minADE and b-minFDE). These metrics are considered more robust for evaluating MP, since they incorporate both the accuracy and confidence of the predicted trajectories. They offer a more comprehensive assessment compared to minADE and minFDE.

LFF-V2V results for object detection are shown in Table I. Here, we can see that WBF performs better than NMS in ROS-based communications, where there are no latencies. However, when time-loss degradation introduces noise to the system, like in V2X communications, WBF's fusion mechanism is negatively impacted as shown in the following results. Despite V2X performance drop compared to ROS, NMS still improves AP@0.7 in 2.37 % for CAV 1 and 2.73 % for CAV 2 respectively. On the other hand, WBF declines 2.7% for CAV 1 and 1.8% for CAV 2 in AP@0.7 under V2X conditions. These drops illustrate the difficulty WBF faces when handling incomplete or delayed data. Furthermore, because WBF combines overlapping boxes instead of

TABLE I: LFF-V2V Average Precision performance in CARLA. We show the CAVs, communication context, late fusion methods and outcome metrics. The ”-” denotes that there is no communication nor fusion method used. Total percentage of the base detector and after applying each method is represented in brackets.

CAV	Comm	Method	AP@0.5 ↑	AP@0.7 ↑
1	-	-	73.7 (100.0)	70.4 (100.0)
	ROS	WBF	**79.4 (108.5)**	**74.9 (106.4)**
		NMS	78.7 (106.7)	72.3 (102.7)
	V2X	WBF	73.2 (99.3)	68.5 (97.3)
		NMS	**76.1 (103.4)**	**72.1 (102.4)**
2	-	-	72.5 (100.0)	69.2 (100.0)
	ROS	WBF	**79.0 (108.9)**	**72.9 (105.4)**
		NMS	77.1 (106.5)	71.3 (103.1)
	V2X	WBF	72.8 (100.5)	67.9 (98.2)
		NMS	**74.1 (103.3)**	**71.8 (102.7)**

TABLE II: LFF-V2V Motion Prediction performance in CARLA. We show the CAVs, communication context, late fusion methods, and outcome metrics. The ”-” denotes that there is no communication nor fusion method used. Total percentage of the base prediction and after applying each method is represented in brackets.

CAV	Comm	Method	minADE (m) ↓	minFDE (m) ↓	b-minADE (m) ↓	b-minFDE (m) ↓
1	-	-	1.08 (100.0)	1.61 (100.0)	1.70 (100.0)	2.23 (100.0)
	ROS	WBF	0.92 (114.8)	1.41 (112.4)	1.44 (115.3)	2.03 (109.0)
		NMS	**0.79 (126.8)**	**1.27 (121.1)**	**1.41 (117.1)**	**1.88 (115.7)**
	V2X	WBF	1.15 (93.5)	1.72 (93.2)	1.85 (91.2)	2.45 (90.2)
		NMS	**0.98 (109.3)**	**1.45 (110.0)**	**1.60 (105.8)**	**2.10 (105.8)**
2	-	-	1.15 (100.0)	1.62 (100.0)	1.79 (100.0)	2.26 (100.0)
	ROS	WBF	1.00 (113.4)	1.34 (117.3)	1.43 (120.1)	1.96 (113.3)
		NMS	**0.78 (132.2)**	**1.24 (123.4)**	**1.39 (122.3)**	**1.85 (118.1)**
	V2X	WBF	1.25 (91.3)	1.67 (96.7)	1.90 (93.7)	2.50 (89.4)
		NMS	**1.05 (108.7)**	**1.42 (112.3)**	**1.60 (110.6)**	**2.05 (109.3)**

discarding them, like NMS, it fails to eliminate redundancy. When introducing delays, WBF can compound errors and put noise into the fused outputs.

The difference in how these methods handle redundancy has significant implications for MP. WBF's inability to eliminate redundancy leads to higher trajectory prediction errors than NMS, as shown in Table II. We can see here that, even in ROS communications, NMS outperforms WBF, obtaining 17.1 % and 15.7 % decrease for CAV 1 and 22.3 % and 18.1 % decrease for CAV 2 in b-minADE and b-minFDE, respectively. V2X degradation also affects MP metrics, but as in object detection, NMS still improves the overall system, whereas WBF struggles. We can see how WBF underperforms with a 8.8% and 9.8% increase for CAV 1 and 6.3% and 10.6% increase for CAV 2 in b-minADE and b-minFDE. In comparison, NMS provides a 5.8% decrease in CAV 1 and 10.6% and 9.3% decrease for CAV 2 for these metrics in the same setup.

Figure 3 shows qualitatively these performance differences between NMS and WBF in ROS communications. Here, we can appreciate that CAV 1 vision helps CAV 2 to overcome occlusions due to its surrounding vehicles, leading to better forecasts.

B. V2X-Real-V2V

In V2X-Real-V2V, we apply the same considerations for MP metrics mentioned in CARLA evaluation. LFF-V2V results for MP are shown in Table III. From these results we can conclude that NMS still improves MP, even in degraded V2X conditions, with decreases of 5.8% and 4.2% for CAV 1 and 6.1% and 4.5% for CAV 2 in b-minADE and b-minFDE. However, we can see quite a remarkable performance drop compared to ROS conditions, where there are decreases of 15.2% and 18.3 % for CAV 1 and 11.2% and 15.5% for CAV 2 in these metrics.

TABLE III: LFF-V2V Motion Prediction performance in V2X-Real-V2V. We show the CAVs, communication context, late fusion methods, and outcome metrics. The ”-” denotes that there is no communication nor fusion method used. Total percentage of the base prediction and after applying NMS is represented in brackets.

CAV	Comm	minADE (m) ↓	minFDE (m) ↓	b-minADE (m) ↓	b-minFDE (m) ↓
1	-	0.33 (100.0)	0.44 (100.0)	1.38 (100.0)	1.20 (100.0)
	ROS	**0.21 (136.4)**	**0.27 (138.6)**	**1.17 (115.2)**	**0.98 (118.3)**
	V2X	**0.28 (115.2)**	**0.37 (115.9)**	**1.45 (105.8)**	**1.25 (114.2)**
2	-	0.45 (100.0)	0.47 (100.0)	0.98 (100.0)	1.10 (100.0)
	ROS	**0.28 (137.8)**	**0.23 (151.1)**	**0.87 (111.2)**	**0.93 (115.5)**
	V2X	**0.35 (122.2)**	**0.39 (117.0)**	**0.92 (106.1)**	**1.05 (104.5)**

LFF-V2V qualitative results in ROS-based setup are shown in Figure 4, which illustrates CAV 2 exiting an intersection. It gets an enhanced prediction of the following agents thanks to fused detections with CAV 1.

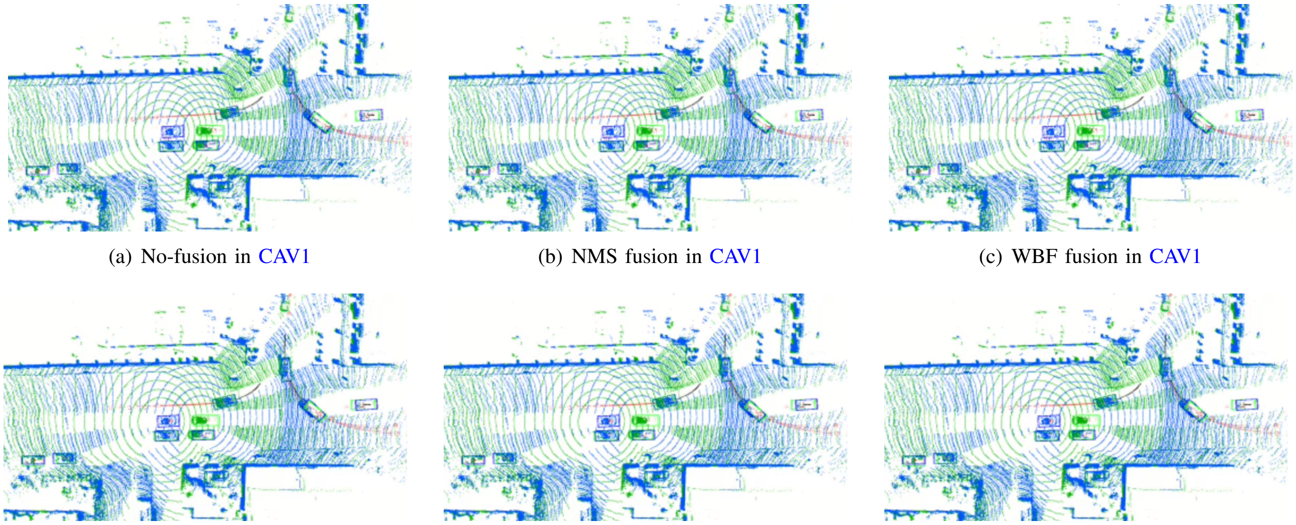
VI. CONCLUSIONS AND FUTURE WORKS

Our research presents LFF-V2V: A Late Fusion Cooperative Framework in V2V Scenarios. For this purpose, we have implemented two SOTA late fusion methods, Non Maximum Suppresion [12] and Weighted Box Fusion [13], and evaluate their performance under different communication modes: ROS and V2X. This evaluation gathers V2V object detection and motion prediction analysis in CARLA, and trajectory forecasting in a real V2X dataset (V2X-Real [16]). For object detection, we have trained PointPillars [33] model. For motion prediction, we have used a SOTA single-vehicle map-less model (HiVT) as in our previous work [17]. Our results, based on SOTA metrics, prove that our CP framework achieves a more precise scene context thanks to the vision of other CAVs. The analysis also shows how late fusion techniques are affected by time-loss degradation in object detection and MP.

We intend to transition from simulated V2X environments to real-world scenarios, incorporating actual V2X devices to evaluate the framework under practical conditions. Furthermore, we want to extend the analysis to include mid-fusion techniques and infrastructure-to-vehicle (I2V) setups that will provide a broader understanding of CP in diverse applications.

VII. ACKNOWLEDGEMENTS

This research has been conducted as part of the EVENTS project, which is funded by the European Union, under grant agreement No. 101069614 and within the MEDUSA program under grant number CER-2023101, Red de Excelencia CERVERA which was founded by the MICIN throughout CDTI under the MRR of the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union



(a) No-fusion in CAV1

(b) NMS fusion in CAV1

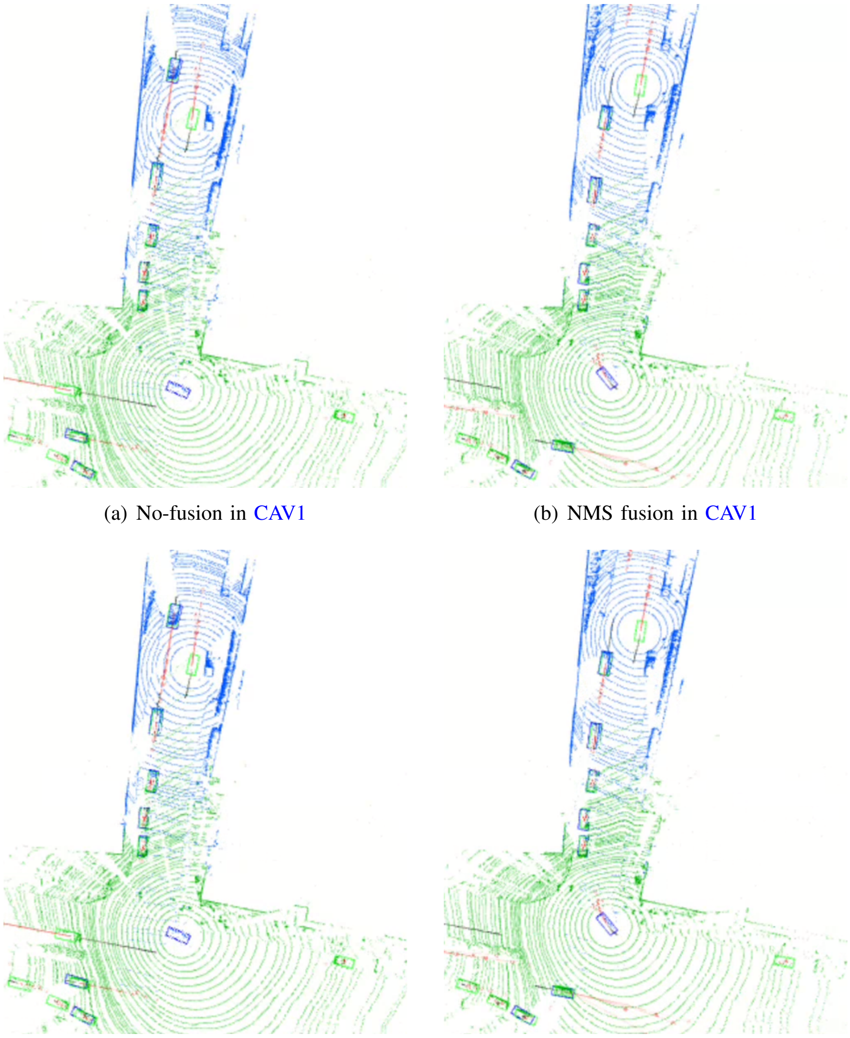
(c) WBF fusion in CAV1

(d) No fusion in CAV2

(e) NMS fusion in CAV2

(f) WBF fusion in CAV2

Fig. 3: CARLA qualitative results. We represent: the CAV1 point cloud, the CAV2 point cloud, the agents detected by CAV1 and CAV2, the **past observations** and our multi-modal prediction. From left to right, we show no fusion, NMS and WBF.



(a) No-fusion in CAV1

(b) NMS fusion in CAV1

(c) No-fusion in CAV2

(d) NMS fusion in CAV2

Fig. 4: V2X-Real-V2V qualitative results. We represent: the CAV1 point cloud, the CAV2 point cloud, the agents detected by CAV1 and CAV2, the **past observations** and our multi-modal prediction. From left to right, we show no-fusion and NMS.

or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," 2023.

[2] M. Gulzar, Y. Muhammad, and N. Muhammad, "A survey on motion prediction of pedestrians and vehicles for autonomous driving," *IEEE Access*, vol. 9, pp. 137957–137969, 2021.

[3] H. Ruan, H. Yu, W.-Y. Yang, S. Fan, Y. Tang, and Z. Nie, "Learning cooperative trajectory representations for motion forecasting," *ArXiv*, vol. abs/2311.00371, 2023.

[4] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, K. Oguchi, and Z. Huang, "A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation," 2022.

[5] Cooperative Driving Automation(CDA) Committee, *Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles*, may 2020.

[6] H. Yu, W. Yang, J. Zhong, Z. Yang, S. Fan, P. Luo, and Z. Nie, "End-to-end autonomous driving through v2x cooperation," 2024.

[7] 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*, vol. 25, no. 9, pp. 12163–12175, 2024.

[8] M. Yazgan, T. Graf, M. Liu, T. Fleck, and J. M. Zöllner, "A survey on intermediate fusion methods for collaborative perception categorized by real world challenges," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2226–2233, 2024.

[9] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," 2020.

[10] Y. Lu, Y. Hu, Y. Zhong, D. Wang, Y. Wang, and S. Chen, "An extensible framework for open heterogeneous collaborative perception," in *The Twelfth International Conference on Learning Representations*, 2024.

[11] Cooperative Driving Automation(CDA) Committee, *Process for Developing an Interoperable Cooperative Driving Use Case Test Framework and Test Procedures*, aug 2023.

[12] J. H. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," *CoRR*, vol. abs/1705.02950, 2017.

[13] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104117, 2021.

[14] Z. Zhou, L. Ye, J. Wang, K. Wu, and K. Lu, "Hivt: Hierarchical vector transformer for multi-agent motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," 2017.

[16] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng, L. Jin, M. Lei, Z. Ma, Z. He, H. Ma, Y. Yuan, Y. Zhao, and J. Ma, "V2x-real: A.largs-scale dataset for.vehicle-to-everything cooperative perception," in *Computer Vision – ECCV 2024* (A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds.), (Cham), pp. 455–470, Springer Nature Switzerland, 2025.

[17] J. Araluce, A. Justo, A. Arizala, L. González, and S. Díaz, "Enhancing motion prediction by a cooperative framework," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1389–1394, 2024.

[18] T. Lyu, M. Noor-A-Rahim, D. Pesch, and A. O'Driscoll, "A survey and tutorial of redundancy mitigation for vehicular cooperative perception: Standards, strategies and open issues," 2025.

[19] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22668–22677, 2024.

[20] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," 2022.

[21] 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," 2022.

[22] M. Yazgan, M. V. Akkanapragada, and J. M. Zoellner, "Collaborative perception datasets in autonomous driving: A survey," 2024.

[23] G. Kueppers, J.-P. Busch, L. Reiher, and L. Eckstein, "V2aix: A multi-modal real-world dataset of etsi its v2x messages in public road traffic," 2024.

[24] X. Zhang, S.-Y. Cao, F. Wang, R. Zhang, Z. Wu, X. Zhang, X. Bai, and H.-L. Shen, "Rethinking early-fusion strategies for improved multispectral object detection," *IEEE Transactions on Intelligent Vehicles*, pp. 1–15, 2024.

[25] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," 2022.

[26] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," 2022.

[27] T. D. Borba, O. Vaculín, H. Marzbani, and R. N. Jazar, "Increasing safety of automated driving by infrastructure-based sensors," *IEEE Access*, vol. 11, pp. 94974–94991, 2023.

[28] J. Wang, G. Ren, F. Zhou, H. Zhang, G. Luo, Q. Yuan, and J. Li, "Practical collaborative perception: Design, implementation, and evaluation," in *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pp. 1–6, 2023.

[29] S. Ochs, M. Yazgan, R. Polley, A. Schotschneider, S. Orf, M. Uecker, M. Zipfl, J. Burger, A. Vivekanandan, J. Amritzer, M. R. Zofka, and J. M. Zöllner, "Empowering autonomous shuttles with next-generation infrastructure," 2024.

[30] J. Schmidt, J. Jordan, F. Gritschneder, and K. Dietmayer, "Crat-pred: Vehicle trajectory prediction with crystal graph convolutional neural networks and multi-head self-attention," 2022.

[31] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," 2023.

[32] Z. Wang, Y. Wang, Z. Wu, H. Ma, Z. Li, H. Qiu, and J. Li, "Cmp: Cooperative motion prediction with multi-agent communication," *arXiv preprint arXiv:2403.17916*, 2024.

[33] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12689–12697, 2019.

[34] T. Sun, M. Segu, J. Postels, Y. Wang, L. V. Gool, B. Schiele, F. Tombari, and F. Yu, "Shift: A synthetic driving dataset for continuous multi-task domain adaptation," 2022.

[35] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sept. 2016.

[36] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, May 2022.

[37] M.-F. Chang, J. W. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] A. Justo, J. Araluce, J. Romera, M. Rodriguez-Arozamena, L. González, and S. Díaz, "Simbusters: Bridging simulation gaps in intelligent vehicles perception," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2471–2476, 2024.