Enhancing Motion Prediction by a Cooperative Framework

Javier Araluce[1], Alberto Justo[1], Asier Arizala[1,2], Leonardo González[1] and Sergio Díaz[1]

***Abstract*— Cooperative perception is a technique that enhances the on-board sensing and perception of automated vehicles by fusing data from multiple sources, such as other vehicles, roadside infrastructure, cloud/edge servers, among others. It can improve the performance of automated driving in complex scenarios, like unsignalled roundabouts or intersections where the visibility and awareness of other road users are limited. Motion Prediction (MP) is a key component of cooperative perception, as it enables the estimation and prediction of microscopic traffic states, such as the positions and speeds of all vehicles. It relies on information from other agents and their relationships among them, so the information provided by external sources is valuable because it enhances the understanding of the scene.**

In this paper, we present improved MP through Vehicle to Vehicle (V2V) communication. We have trained Hierarchical Vector Transformer (HiVT) to be a map-less solution that can be used in road domains. With this model, we have implemented and compared two association methods to evaluate our framework on a real V2V dataset (V2V4Real). Our evaluation concludes that our V2V MP improves performance due to better scene understanding over a single-vehicle MP.

I. Introduction

Perception systems have been a crucial development task for self-driving vehicles. The recent developments use deep learning for different perception tasks such as 3D object detection [1], [2], object tracking [3], [4], semantic segmentation [5], [6] and motion prediction [7], [8]. However, these systems do not take advantage of the multi-view provided by a collaborative framework. These systems suffer from several challenges, such as occlusion and near vision, which limit their performance [9]. A cooperative framework is particularly suitable to improve a motion prediction system because the information needed to process the data (agent positions) is light (due to limited bandwidth). In addition, consideration of social interaction can take advantage of more agents in the scene to make a more accurate prediction.

Motion prediction, also known as motion forecasting, addresses the challenge of predicting future trajectories of dynamic agents surrounding the ego-vehicle. Predicting the future behaviour of traffic agents around the ego-vehicle is one of the key ongoing challenges in reaching full self-driving autonomy [10], [11]. Predicted trajectories help the motion planning system to achieve an efficient and safe path for the ego-vehicle. Autonomous vehicles need to understand the environment to anticipate the future on the road ahead.

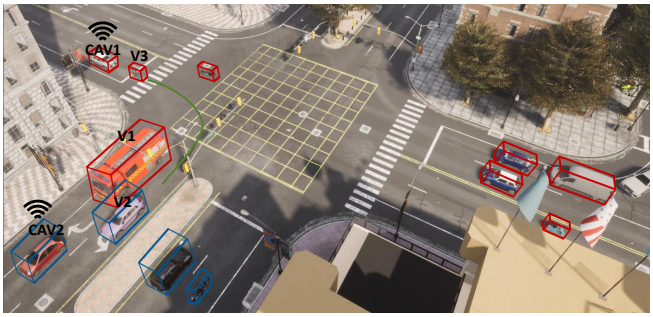


Fig. 1: Illustration of CP to accurately predict the movement of surrounding agents. Red bounding boxes are detected by CAV 1 and blue bounding boxes are only detected by CAV 2. The predicted trajectory of an agent is shown in green.

Despite their advancements, single-vehicle perception technologies continue to face significant hurdles. The primary challenges include sparse observations due to occlusion, limited sensor Field of View (FoV), and sensor resolution [12]. Furthermore, the system's robustness is questioned due to its vulnerability to sensor errors caused by adverse weather conditions and hardware failures [13]. These issues limit the potential and safety of automated vehicles in intricate driving situations.

The field has seen the emergence of Cooperative Perception (CP) technology to mitigate these limitations. This technology emphasizes the interaction between multiple agents to compensate for the shortcomings in single-vehicle perception. The essence of CP is the fusion of sensor data from various agents [14].

Figure 1 shows an example of an intersection where CP is crucial for the prediction of the movement of the surrounding agents. Vehicle 2 is occluded by Vehicle 1, a fire truck, and CAV 1 cannot accurately predict the movement of Vehicle 3, a motorcycle, as a result. However CAV 2 can see vehicle 2 and transmits this information to CAV1 for a better trajectory prediction. Only one agent trajectory has been represented for a clearer explanation.

The following paper presents a study of cooperative perception focused on motion prediction. The analysis was performed on a real Vehicle to Vehicle (V2V) dataset (V2V4Real [15]) using a modified state-of-the-art motion prediction model [7]. We used two association methods to avoid double detections. The contributions are as follows:

- We have adapted and trained HiVT [7] in Argoverse 1 [16] to be a map-less solution model that we can use in other domains where the map is not available.
- We have studied two association methods for CP: Eu-

[1]Javier Araluce, Alberto Justo, Asier Arizala, Leonardo González and Sergio Díaz are with TECNALIA, Basque Research and Technology Alliance (BRTA), 48160 Derio, Bizkaia, Spain javier.araluce, alberto.justo, asier.arizala, leonardo.gonzalez, sergio.diaz {@tecnalia.com}

[2]Asier Arizala is Department of Automatic Control and Systems Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain

clidean distance to the connected vehicle and bounding box clustering.
- We have evaluated the framework performance in a novel V2V real dataset (V2V4Real [15]) with an extensive comparison of different options.

II. Related works

A. Cooperative perception for Motion Prediction

Predicting the trajectory of surrounding dynamic objects has been widely studied in the literature. The development has been based on some publicly available datasets. Some of these datasets are: NuScenes Prediction [17] is a publicly available large-scale dataset consisting of 40k scenarios for AD. The Waymo Open Motion Prediction [18] which has 1.1 M examples. Argoverse 1 Motion Forecasting Dataset [16], which is composed of 324,557 trajectories.

However, while these datasets have taken the motion prediction task to the next level, they do not consider information from associated perceptual providers. To address this challenge, several datasets have been collected that capture perceptual systems from different perspectives. Table I shows some of these datasets.

Some of these datasets have been collected in simulation environments [19]–[21]. They have used the CARLA simulator [22] to collect data from different perspectives (vehicles and infrastructure). However, although simulation environments allow unlimited data collection, they are biased by over-simplified vehicle physics. In particular, motion prediction requires different driving behaviours that are not biased by the simulator physics. For this reason, other datasets have been collected in real data [15], [23]–[25].

DAIR-V2X [24] was the first large-scale vehicle-infrastructure cooperative autonomous driving dataset. It collects data from the infrastructure (10k frames) and a vehicle (22k frames). Incidentally, the data must be approved to be downloaded outside of China.

The A9 dataset [23] collected 4.8k frames from two cameras and two LiDARs placed on infrastructure. They do not provide information from a connected vehicle. Since motion prediction is mainly needed for the motion planning of a vehicle, we believe that the information should come from at least one CAV. Moreover, they plan to address this concern in the near future.

Recently, the V2X-Seq dataset [25] was released. It is the first large-scale sequential V2X dataset. The temporal information makes it perfect for motion prediction. However, as DAIR-V2X, its access outside China is restricted, so we were not able to use it in this work.

For these reasons, we have used V2V4Real [15], which consists of data from two vehicles recorded in real environment. They do not provide sequence information, but the dataset is prepared for tracking purposes, so we adapted the dataset to provide sequence information. We will provide detailed information about the dataset in future sections of this work.

B. Motion prediction

There are several approaches within the motion prediction paradigm to solve this complex challenge. Most of them rely on the map to increase their accuracy [7], [10], [26]. It supplements the online information with high-fidelity maps. However, this information is not always available and makes the model less scalable to other domains, as the construction of HD maps is expensive and time-consuming [27].

Nonetheless, other models do not require a map and are, therefore, suitable for this work. CRAT-Pred [28] uses a graph convolution method to obtain the trajectory predictions. However, its output lacks confidence for the k-predictions (k = 6) and is not fast enough for real-time applications. The authors in [27] propose a map-free method that gives on-pair results on accuracy with other SOTA methods based on maps. However, they have not published their code, and their results are not reproducible. Nevertheless, we have based our work on HiVT [7], which won the Argoverse 1 challenge in 2022. The original model uses map information that is encoded to produce high-fidelity trajectories.

III. Framework

This section explains the construction of the collaborative framework developed (Figure 2). First, the dataset is modified for use in the MP task. Then, the HiVT architecture is transformed to be used without a map. Finally, the association methods are used to filter the same detections from two sources.

A. Dataset

The V2V4Real [15] dataset has three perception tasks: cooperative 3D object detection, cooperative 3D object tracking, and Sim2Real domain adaptation for CP. Nevertheless, we aim to prove its value for motion prediction, which is another perception task that can be enhanced thanks to collaborative information.

Following the method proposed in Argoverse 1 [16], a motion prediction dataset was constructed using the information from the two vehicles of V2V4Real. First, we obtain the detections from both cars. To do this, we used the ground truth of the dataset as we did not want to introduce noise from a detector into our method. Then, we transform all the detections into the same coordinate system using the transformation matrices provided by the dataset (Equation 1). P is the local point, T is the transformation matrix and P' is the global point. Furthermore, we store their local positions with the CAV source as a coordinate frame. We also collect and transform the point clouds for visualisation and the association method. As we carry out the prediction in Bird's Eye View (BEV), we have disregarded the z coordinate.

$$P' = T \cdot P \tag{1}$$

For the temporal information, we used 50 frames recorded at 10 Hz. Following the same configuration as Argoverse, where the past data is 2 seconds, and the predicted horizon is 3 seconds. We feed the past information into the network

TABLE I: Comparison between Cooperative Perception-related Datasets.

Dataset	Real/Sim	V2X	Size (km)	Lidar pcds	Maps	3D boxes	Classes	Locations
OPV2V [19]	Sim	V2V	-	11k	Yes	230k	1	CARLA
V2X-Sim [20]	Sim	V2V&I	-	10k	Yes	26.6k	1	CARLA
V2XSet [21]	Sim	V2V&I	-	11k	Yes	230k	1	CARLA
A9 Intersection [23]	Real	V2I	-	4.8k	No	57.4k	10	Hanover, Germany
DAIR-V2X [24]	Real	V2I	20	39k	No	464k	10	Beijing, CN
V2X-Seq [25]	Real	V2V&I	-	210k (seq)	No	20,301k (2D)	8	Beijing, CN
V2V4Real [15]	Real	V2V	410	20k	Yes*	240k	5	Ohio, USA

Notes: * indicates that the map are listed as public but they have not been released by the day of this work.

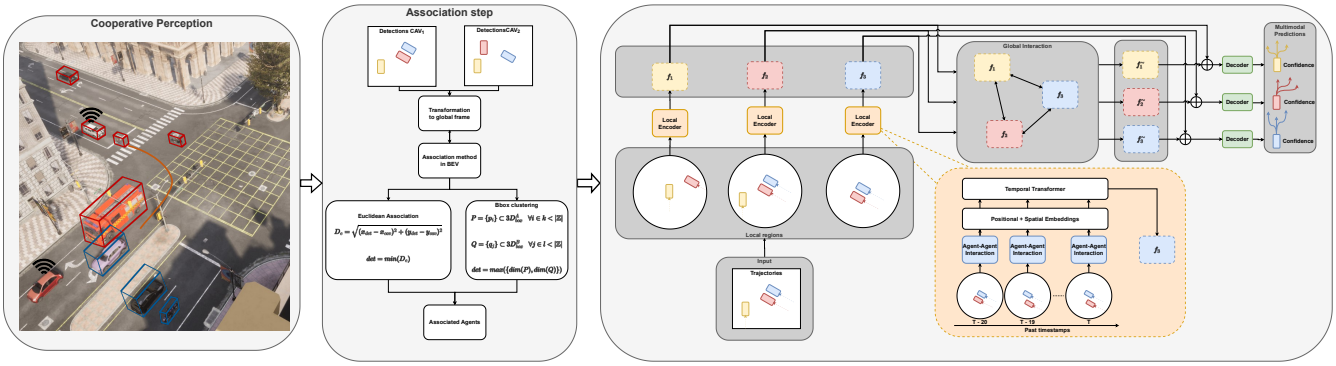


Fig. 2: Cooperative Framework for Motion Prediction

TABLE II: Performance of the motion prediction models in Argoverse 1 validation set.

Model	Map	minADE ↓	minFDE ↓	MR (%) ↓
HiVT-64 [7]	✓	0.69	1.04	0.10
HiVT-128 [7]	✓	**0.66**	**0.96**	**0.09**
Crat-Pred [28]	✗	0.85	1.44	0.17
HiVT-64 (ours)	✗	0.76	1.24	0.14

and use the future one to compare with the predictions to get performance metrics on the method.

B. HiVT

To obtain the predictions, we based our proposal on HiVT [7] with the map information disabled. We train the model in Argoverse 1 to learn the spatio-temporal relationships between the agents in the scene. We first organise a traffic scene as a collection of vectorised entities. The model consists of a local encoder that encodes the agent-agent interactions. This representation is rotation-invariant for each agent. A global module encodes the long-range dependencies between the local agent representations, which are agent-centric. Finally, a multimodal future decoder that predicts all agent trajectories in a single pass. We represent the output in the local coordinate frame, the Autonomous Vehicle (AV) position, which is one of the CAVs. We will evaluate changing this local framework to measure the impact of the viewpoint in our collaborative framework.

Table II compares the obtained results in Argoverse 1 by different models, evaluated in terms of three metrics used in motion prediction: Minimum Average Displacement Error (minADE), Minimum Final Displacement Error (minFDE) and Miss Rate (MR).

HiVT has been presented, using map information, in two modalities (embedded dimension of 64 and 128). The best performance being achieved with an embedded dimension of 128. The modification made to HiVT to operate without a map (embedded dimension of 64), as per our current proposal, results in a 10% decrease in minADE, a 19% decrease in minFDE, and a 4% decrease in MR performance as the cost for allowing map free application in different locations.

C. Association methods

We compared two methods of association in the presence of multiple perception systems. It is important to note that these methods do not consider the confidence of detections, as we use the ground truth from the dataset. In collaborative perception systems where the confidence of the detectors is available, alternative methods should be used. The two CAVs from the scene are denoted by A and B.

1) Association by Euclidean distance: The study assessed the correlation between objects with the same tracker in the collaborative framework. We select the detection closest to the CAV for greater accuracy. Equation 2 shows the Euclidean distance calculation. We represent each detection by its position in BEV ($Det_{pos} = x_{det}, y_{det}$). The Connected Autonomous Vehicle (CAV) position is also in BEV ($CAV_{pos} = x_{CAV}, y_{CAV}$). We then select the detection with the minimum distance for a pair of detections that share the same tracker, as shown in Equation 3.

$$D_e = \sqrt{(x_{det} - x_{cav})^2 + (y_{det} - y_{cav})^2} \tag{2}$$

$$det = argmin(D_e^A, D_e^B) \tag{3}$$

2) Association by bounding box clustering: We tested an alternative association method in addition to the Euclidean distance to the CAV. This method measures the number of points from the point cloud within the bounding box. It provides higher accuracy than the Euclidean distance, as it considers occlusions.

The Equation 4 is the set (P) of points (p_i) from the point cloud having k samples that fall within the bounding box ($3D^A_{box}$). The Equation 5 does the same calculation for the point cloud with l samples falling within the bounding box ($3D^B_{box}$). The $3D^A_{box}$ and $3D^B_{box}$ bounding boxes refer to the same object. We then compare both sets (P, Q) to select the larger one (Equation 6), which will be our detection.

$$P = \{p_i\} \subset 3D^A_{box} \quad \forall i \in k < |\mathbb{Z}| \tag{4}$$

$$Q = \{q_j\} \subset 3D^B_{box} \quad \forall j \in l < |\mathbb{Z}| \tag{5}$$

$$det = max(\{dim(P), dim(Q)\}) \tag{6}$$

IV. Experimental setup

We used the complete V2V4Real dataset, consisting of 20,000 frames, for our experiment. We created a sequenced dataset with it to evaluate motion prediction. The original authors divided into three days the dataset, with several sequences per day containing the recorded frames. The first day has 2,726 sequences, the second has 3,246, and the third has 1,644. We evaluated a total of 7,616 over three days. It is not a large dataset for motion prediction, so we did not use it for training and only evaluated a trained model. The model used was trained in Argoverese 1, as mentioned above.

We used common metrics for evaluating motion prediction [7], [16]: Minimum Average Displacement Error (minADE), Minimum Final Displacement Error (minFDE), Miss Rate (MR), Percentage of minADE (p-minADE), Percentage of minFDE (p-minFDE), Percentage of Miss Rate (p-MR), Brier Score for minADE (brier-minADE) and Brier Score for minFDE (brier-minFDE).

V. Results

In this section, we describe the results of our experiment. We evaluated eight different V2V + association combinations:

- We tested the two CAVs on their own (Tesla and Astuff), without V2V enhanced perception.
- We evaluated the enhanced motion prediction by V2V, first without any association method and then with the Euclidean and bbox clustering association methods, to get a better surround understanding.
- We evaluated changing the viewpoint for the motion prediction section between the two possible CAVs, following the work proposed in [29]. It determines which CAV is at the center of the scene. Making all the trajectories related to this frame.

Table III shows the obtained results for the main motion prediction metrics selected above. We can observe the performance of the single-vehicle method is worse compared to the V2V enhanced method. Additionally, we present the average number of actors. We conclude that since there are more actors in the V2V scene, it increases the overall error but gives more information about the environment. For this reason, we have normalised all metrics by the number of actors. These metrics present a better conclusion about the surround understanding. These results can be seen in Table IV.

Furthermore, the results of the two association methods do not allow us to conclude which one is better, as there is not a marked difference between the two results. However, we can see a slight difference for all the FDE metrics, where the Euclidean association gives a better result (Table III).

When changing the point of view of the model, we observe the same behaviour. There is not significant difference between one system and the other. This is due to HiVT being translation and rotation invariant.

As mentioned above, Table IV shows the comparison between the different methods, but considering the number of actors. This comparison gives a better conclusion about our work, which improves the motion prediction performance by increasing the number of actors. For better visualisation, we have presented the calculated performance increase in percentage in Table V. We can see a better performance thanks to our improved MP, with an increase in all metrics between 13 % and 24 % over the single-vehicle method. Moreover, it shows a better increase in performance in the "brier" metrics, which leads to the conclusion that the model not only predicts better the k-options but also the confidence for each possible trajectory.

Finally, we have shown some qualitative results from a sequence we have evaluated (Figure 3). Figures 3(a) and 3(d) show the CAV without enhanced MP. Figures 3(b) and 3(e) show the MP with V2V using the Euclidean association and Figures 3(c) and 3(f) with the bbox clustering association. We can see an improvement between figures 3(c) and 3(e). They show that the information from the CAV ahead helps to get a better trajectory prediction. The reason for this is that the CAV behind is unable to see the truck ahead or the vehicles on the side.

VI. Conclusion and Future works

We have presented our Enhanced Motion Prediction by a Cooperative Framework. For this purpose, we have adapted and trained a SOTA model (HiVT) from a single vehicle MP and evaluated it on a real V2V dataset (V2V4Real [15]). In addition, we have used two association methods that do not require the confidence of the detectors, as we have used the GT from the dataset. We have evaluated a Euclidean and a Bounding Box clustering association method. We have presented our results using SOTA metrics and demonstrated that our collaborative framework achieves a better scene understanding thanks to the information provided by other CAVs. It can be inferred that the results can be enhanced

TABLE III: Comparison of methods on the V2V4Real dataset. We show the CAVs, the association method, the viewpoint, the number of actors considered and the performance metrics. The "-" denotes that there is no association method used.

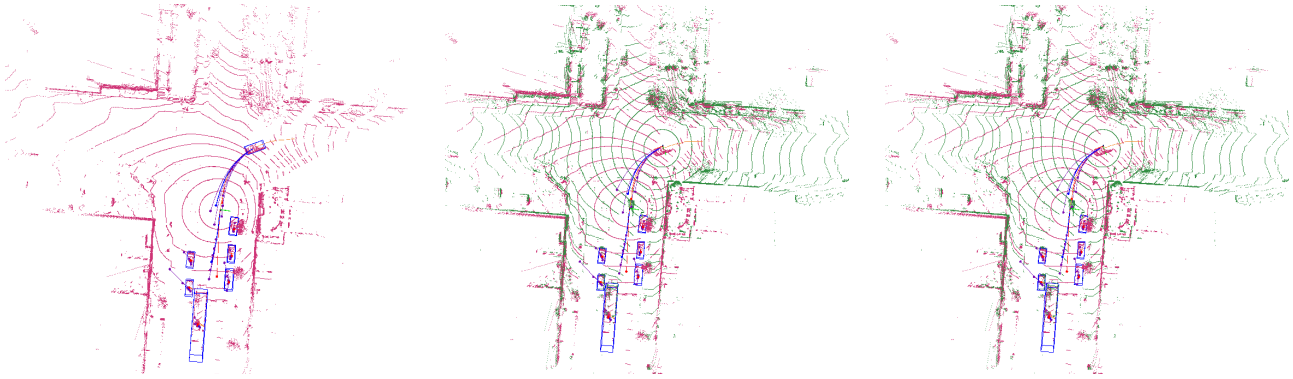
CAVs	Association	Viewpoint	Number Actors	minADE (m) ↓	minFDE (m) ↓	MR ↓	p-minADE ↓	p-minFDE ↓	p-MR ↓	brier-minADE (m) ↓	brier-minFDE (m) ↓
Tesla	-	Tesla	7.74	**1.14**	**2.22**	**0.32**	**2.84**	**3.91**	**0.87**	**1.80**	**2.88**
Astuff	-	Astuff	8.45	1.22	2.30	0.33	2.90	3.99	**0.87**	1.88	2.96
V2V	-	Tesla	14.58	1.34	2.51	0.33	3.02	4.19	**0.87**	2.00	3.17
V2V	Euclidean	Tesla	10.19	1.26	2.37	**0.32**	2.95	4.05	**0.87**	1.92	3.02
V2V	Bbox clustering	Tesla	10.19	1.26	2.37	**0.32**	2.95	4.06	**0.87**	1.92	3.03
V2V	-	Astuff	14.58	1.34	2.52	0.33	3.03	4.21	**0.87**	2.00	3.18
V2V	Euclidean	Astuff	10.19	1.27	2.38	**0.32**	2.95	4.07	**0.87**	1.92	3.04
V2V	Bbox clustering	Astuff	10.19	1.27	2.39	0.33	2.96	4.08	**0.87**	1.93	3.05

TABLE IV: Comparison of methods on the V2V4Real dataset normalised by the number of actors in the scene. We show the CAVs, the association method, the viewpoint and the performance metrics. The "-" denotes that there is no association method used.

CAVs	Association	Viewpoint	minADE ↓	minFDE ↓	MR ↓	p-minADE ↓	p-minFDE ↓	p-MR ↓	brier-minADE ↓	brier-minFDE ↓
Tesla	-	Tesla	0.15	0.29	0.04	0.37	0.51	0.11	0.23	0.37
Astuff	-	Astuff	0.14	0.27	0.04	0.34	0.47	0.10	0.22	0.35
V2V	Euclidean	Tesla	**0.12**	**0.23**	**0.03**	**0.29**	**0.40**	**0.09**	**0.19**	**0.30**
V2V	Bbox clustering	Tesla	**0.12**	**0.23**	**0.03**	**0.29**	**0.40**	**0.09**	**0.19**	**0.30**
V2V	Euclidean	Astuff	0.13	0.24	**0.03**	**0.29**	**0.40**	**0.09**	**0.19**	**0.30**
V2V	Bbox clustering	Astuff	**0.12**	**0.23**	**0.03**	**0.29**	**0.40**	**0.09**	**0.19**	**0.30**

TABLE V: Increased performance through our V2V framework compared to a single vehicle, taking into account the number of vehicles. We show the comparison, the association method, the viewpoint and the performance metrics.

Comparison	Association	Viewpoint	minADE	minFDE	MR	p-minADE	p-minFDE	p-MR	brier-minADE	brier-minFDE
V2V vs Tesla	Euclidean	Tesla	**16%**	**19%**	**22%**	**21%**	**21%**	**24%**	**19%**	**20%**
V2V vs Tesla	Bbox clustering	Tesla	**16%**	**19%**	**22%**	**21%**	**21%**	**24%**	**19%**	**20%**
V2V vs Astuff	Euclidean	Astuff	13%	14%	19%	15%	15%	17%	14%	14%
V2V vs Astuff	Bbox clustering	Astuff	14%	14%	19%	16%	15%	17%	15%	15%



(a) Single-vehicle Tesla

(b) Euclidean association with the Tesla as viewpoint

(c) Bbox clustering association with the Tesla as viewpoint

(d) Single-vehicle Astuff

(e) Euclidean association with the Astuff as viewpoint

(f) Bbox clustering association with the Astuff as viewpoint

Fig. 3: Qualitative results showing our comparison. We represent: the Tesla point cloud, the Astuff point cloud, the agents the past observations, the ground-truth and our multi-modal prediction (with the highest confidence). We show, from left to right, single-vehicle, euclidean and bbox clustering.

by incorporating map information. We intend to continue with this experiment on this dataset, including the map when it becomes available. To measure the impact of map representation in a V2V MP framework.

Besides, we plan to conduct this study in other collaborative domains. We will conduct an experiment in a simulated environment to create complex scenarios. The aim is to assess the effect of detector errors on the system.

ACKNOWLEDGMENT

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 the grant number CER-2023101, Red de Excelencia CERVERA which was founded by the MICIN thorught 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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9098, 2023.

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

[3] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10359–10366, IEEE, 2020.

[4] Z. Zhao, Z. Wu, Y. Zhuang, B. Li, and J. Jia, "Tracking objects as pixel-wise distributions," in *European Conference on Computer Vision*, pp. 76–94, Springer, 2022.

[5] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[6] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13760–13769, 2022.

[7] 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*, pp. 8823–8833, 2022.

[8] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17863–17873, 2023.

[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. Gómez-Huélamo, M. V. Conde, R. Barea, M. Ocaña, and L. M. Bergasa, "Efficient baselines for motion prediction in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[11] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in *Conference on Robot Learning*, pp. 409–418, PMLR, 2021.

[12] L. Li and C. Sun, "Nlos dies twice: Challenges and solutions of v2x for cooperative perception," *arXiv preprint arXiv:2307.06615*, 2023.

[13] K. Ren, Q. Wang, C. Wang, Z. Qin, and X. Lin, "The security of autonomous driving: Threats, defenses, and future directions," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 357–372, 2019.

[14] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "V2x cooperative perception for autonomous driving: Recent advances and challenges," *arXiv preprint arXiv:2310.03525*, 2023.

[15] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13712–13722, 2023.

[16] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8748–8757, 2019.

[17] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

[18] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] 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)*, pp. 2583–2589, IEEE, 2022.

[20] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10914–10921, 2022.

[21] 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*, pp. 107–124, Springer, 2022.

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

[23] W. Zimmer, C. Creß, H. T. Nguyen, and A. C. Knoll, "A9 intersection dataset: All you need for urban 3d camera-lidar roadside perception," *arXiv preprint arXiv:2306.09266*, 2023.

[24] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21361–21370, 2022.

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

[26] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

[27] J. Xiang, J. Zhang, and Z. Nan, "A fast and map-free model for trajectory prediction in traffics," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 873–879, IEEE, 2023.

[28] 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," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7799–7805, IEEE, 2022.

[29] A. Cui, S. Casas, K. Wong, S. Suo, and R. Urtasun, "Gorela: Go relative for viewpoint-invariant motion forecasting," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7801–7807, IEEE, 2023.