https://doi.org/10.2352/EI.2025.37.15.AVM-110

Data optimization strategies for collaborative perception

Besma Abdali [1] ***; Quentin Picard*** [1] ***; Maryem Fadili*** [1]
[1] ***Institut VEDECOM; 23 bis Allée des Marronniers, 78000 Versailles, France***

Abstract

Collaborative perception for autonomous vehicles aims to overcome the limitations of individual perception. Sharing information between multiple agents resolve multiple problems, such as occlusion, sensor range limitations, and blind spots. One of the biggest challenge is to find the right trade-off between perception performance and communication bandwidth. This article proposes a new cooperative perception pipeline based on the Where2comm algorithm with optimization strategies to reduce the amount of transmitted data between several agents. Those strategies involve a data reduction module in the encoder part for efficient selection of the most important features and a new representation of messages to be exchanged in a V2X manner that takes into account a vector of information and its positions instead of a high-dimensional feature map. Our approach is evaluated on two simulated datasets, OPV2V and V2XSet. The accuracy is increased by around 7% with AP@50 on both datasets and the communication volume is reduced by 89.77% and 92.19% on V2XSet and OPV2V respectively.

Introduction

Collaborative perception in a V2X (vehicle-to-everything) manner aims to overcome the challenges of individual perception with a shared perception by integrating information from other agents, such as road infrastructure or other connected vehicles that improves the accuracy of detection and the robustness of the perception system [2]. It solves problems of individual perception, which are often limited by the embedded sensors and face challenges such as occlusions and blind spots [1, 5, 8, 20]. Although collaborative perception overcomes the limitations of individual perception, it introduces a major new challenge, the increase of bandwidth consumption due to the need of frequent and voluminous data communications between vehicles and other agents. [4, 9, 10, 11, 14, 18, 19].

The efficient use of communication resources is essential for collaborative perception. In order to minimize communication costs, a naive solution is the use of late collaboration, where each agent makes their own predictions and shares their outputs with other agents [2]. Late collaboration saves bandwidth but can result in lower perception performance, due to potentially noisy or incomplete individual outputs [2]. Recent works take into account the intermediate collaboration to optimize the perception and communication trade-off. It involves transferring features generated by deep neural networks from other agents to the ego-vehicle, which fuses these features to make predictions. It provides a trade-off between perception performances and bandwidth consumption, but can lead to information loss and redundancy [2].

Previous works make a hypothesis that once two agents collaborate, they are forced to share perceptual information about all spatial areas. This assumption consumes bandwidth because of a large proportion of spatial areas that may contains information that is not relevant to the perception task. Where2comm [4] has proposed a strategy that only transmits messages over constrained spatial areas. While this method reduces the communication costs, it still requires the transmission of high-dimensional feature maps. To address one of the limitations of intermediate collaboration, our main contributions optimize the collaborative exchanges focusing on three key aspects:

- First, a new cooperative perception pipeline based on the Where2comm algorithm with data optimization strategies
- Then, the addition of a data reduction module that impacts the backbone process.
- Finally, an innovative message representation to be transmitted between the connected agents.

This article is structured as follows: First, the related works are described on collaborative perception. Then, the new pipeline with optimization strategies to reduce the amount of data transmitted between the road users is detailed. The applied methodology and experimental results to evaluate the performances on two different datasets are presented. Finally, the findings of this work are detailed and discussed.

Related works

Communication in multi-agent environments is a well-established field where strategies have evolved from predefined protocols to the adoption of machine learning methods to handle complex situations. CommNet [13] and similar works explored continuous communication in multi-agent systems. Vain [3] introduced the use of attention mechanisms for selective merging of information between agents. These studies have mainly dealt with decision-making tasks using reinforcement learning, a necessity in the absence of direct supervision.

Collaborative perception is an emerging field in the development of multi-agent communication systems for perception tasks. It is reinforced by high-quality data and collaborative methods that optimize the trade-off between performance and bandwidth consumption. Works such as V2VNet [14], which uses multi-turn message passing through graphical neural networks and DiscoNet [12], which applies knowledge distillation, have shown significant improvements in perception and prediction performance. V2X-ViT [15] demonstrates the effectiveness of a new multi-agent attention module to integrate heterogeneous information.

Other works such as When2com [10] and Who2comm [11] use an attention mechanism to select information from relevant collaborators. In addition, Where2comm [4] and CoCa3D [6] extend the selection process to incorporate spatial dimensions. How2comm [17] uses a mutual information-aware communication mechanism to preserve informative features, a spatial-channel filtering for efficient sparsification and a flow-guided de-

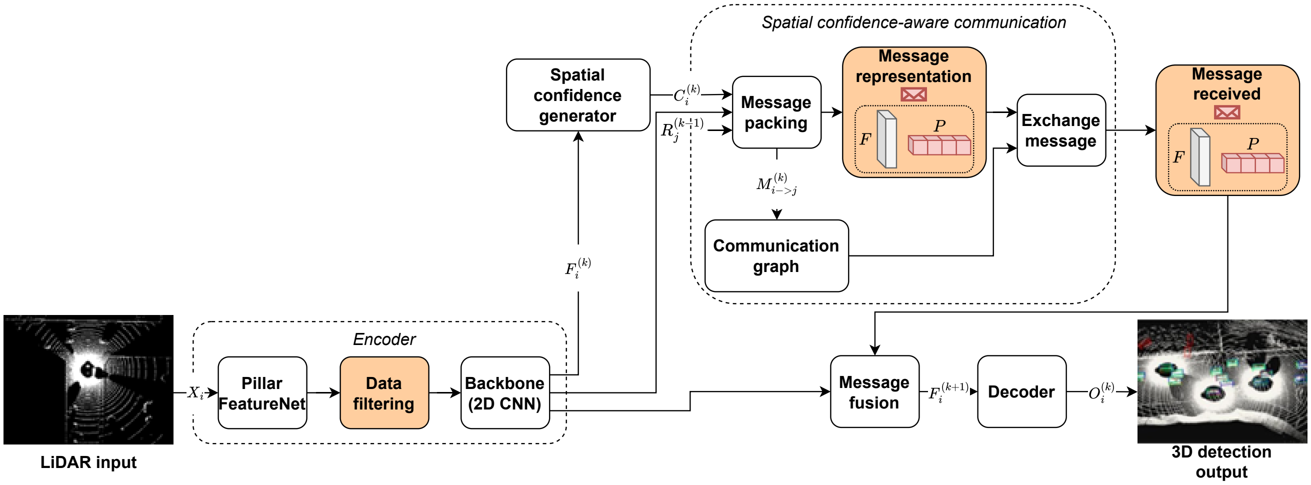

Figure 1. Cooperative perception pipeline based on the Where2comm algorithm with data optimization strategies highlighted in orange. The data filtering module reduces the backbone's dimension and keeps the most important features. The message representation and received modules optimizes the amount of data to be transmitted between multiple road agents.

lay compensation strategy to predict future features and eliminate temporal misalignment. A pragmatic collaboration transformer integrates spatial and contextual semantics among agents. PragComm [7] is based on three main strategies: a pragmatic selection of messages to choose critical data, a representation of messages to allow communication with entire clues and a selection of collaborators to prune unnecessary communication links. It includes two key components: the single-agent detection and tracking and the adaptive pragmatic collaboration to various communication conditions.

Despite these advances, a large part of current methods still transmit superfluous data, resulting in unnecessarily high communication costs. In this study, we seek to optimize the exchange of information by transmitting only features to object recognition, thus reducing the volume of data exchanged and improving the efficiency of the collaborative process. We propose a method where each agent sends a reduced vector of essential information and its spatial position.

Data optimization strategies

This section details the proposed collaborative perception pipeline designed around two main strategies as shown in Figure 1. Our approach is based on the Where2comm algorithm [4], which includes the encoder with the data filtering module. It selects the most relevant and sufficient information for object detection. The second strategy impacts the spatial confidence-aware communication that is responsible for the communication and the fusion of messages from multiple agents. The messages representation module focuses on optimizing the representation of the transmitted messages.

Data Filtering

The data filtering module is located between the Pillar Feature Net (PFN) and the backbone in the encoder part [8]. This module influences the collaborative object detection process from point cloud data. The backbone, which transforms the extracted features by the PFN in an abstract representation, uses a 2D convolutional network to process the pseudo-image computed by the PFN. The data filtering module modifies and reduces the network backbone's dimensions relative to the original input pseudo-image while maintaining essential features. This module influences the spatial confidence-aware message fusion module, where each instance of fused attention is initialized with a feature dimension from the data filtering method. This ensures that the fusion from different agents operates with reduced dimensions feature maps to improve data aggregation.

Several experiments have been carried out to reduce the backbone size by powers of two: $\left[2^{x}, 2^{2x}, 2^{3x}\right]$ where x varies from 0 to 6. The following two configurations were taken into account to achieve an optimal trade-off between accuracy and consumption: [8, 32, 64] and [8, 64, 128].

Message representation

The second optimization strategies is the message representation module to optimize the transmission of the selected feature maps. Each agent employs this message representation based on an information vector **F** which contains the data used for object detection, with its position vector **P**. The approach allows to take into account only non-zeros data. Therefore, it is only necessary to transmit the information and position vector, rather than the entire feature map consisting of floating-point numbers and zeros.

Experimental results

In this section, we present the experimental results. First, we describe the evaluation method and then we compare the performances of our approach with the baseline Where2comm and two other fusion methods, early and late fusion. Early fusion aggregates raw data from different agents while late fusion fuses the 3D bounding boxes predictions computed by each connected agents.

Methods

Our experiments cover two simulated datasets with point cloud processing: V2XSet [15] and OPV2V [16]. The cooperative 3D object detection task is performed with the use of 2 to 7

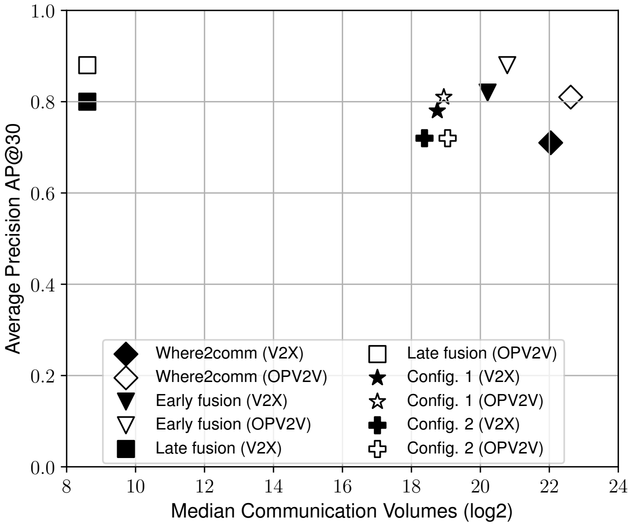

Figure 2. *Average precision at 0.30 IoU threshold (AP@30) in relation to the volume of communication from equation 1.*

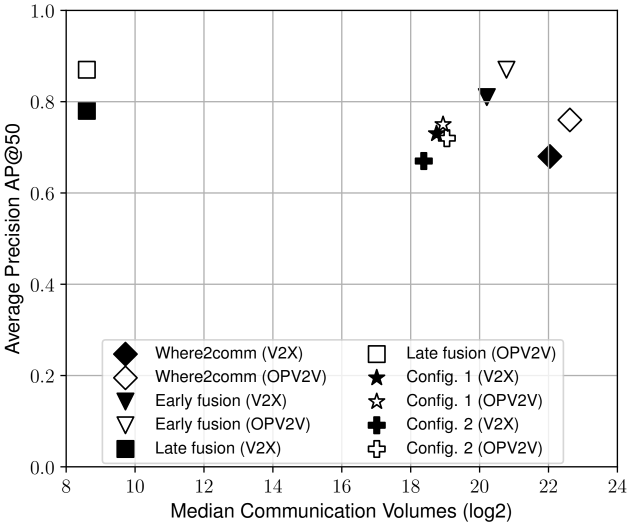

Figure 3. *Average precision at 0.50 IoU threshold (AP@50) in relation to the volume of communication from equation 1.*

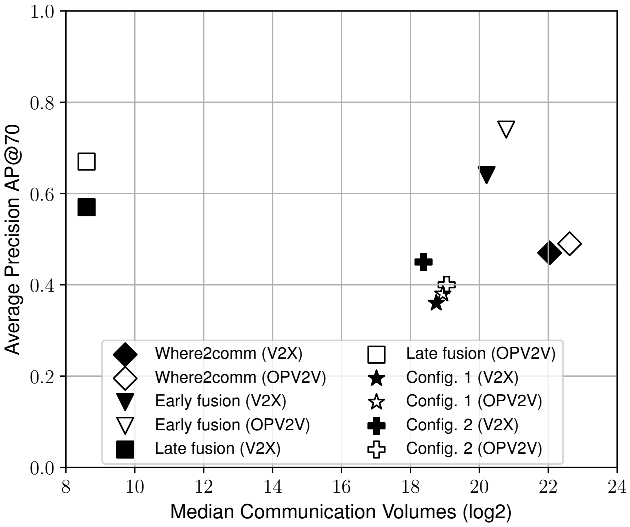

Figure 4. *Average precision at 0.70 IoU threshold (AP@70) in relation to the volume of communication from equation 1.*

connected agents. They are characterized as connected vehicles and infrastructure.

The results are evaluated using three metrics. First, we compute the traditional Average Precision (AP) at three Intersection over Union (IoU) thresholds : 0.30, 0.50 and 0.70. Then we evaluate the communication volume [4] which counts the size of the message per byte in logarithmic scale with base 2 as described in equation 1.

$$\log_2\left(M_{i,j}^{(k)} \times D \times \frac{32}{8}\right) \tag{1}$$

$M_{i,j}^{(k)}$ is the selection matrix, D denotes the channel dimension, 32 is multiplied as float32 data type is used to represent each number, 8 is divided as the metric byte is used.

The final evaluated metric is the execution time in milliseconds (ms) for each part of the pipeline.

Training Details: the model has been trained for 5 epochs with a learning rate of 10^{-3} on NVIDIA GeForce RTX 2080 Ti with the OpenCOOD framework [16]. The basic settings of the backbone are used in the encoder block which corresponds to the size of filters and is fixed to [64,128,256]. This setting is changed in our experiments according to both configurations used.

Performance Analysis

To evaluate the impact of our approach on object detection accuracy and bandwidth consumption, three algorithms have been identified and selected: early fusion collaboration [2], late fusion collaboration [2], and Where2comm [4]. Those algorithms has been selected based on the best trade-off between detection accuracy and bandwidth consumption.

Figures 2, 3 and 4 show the average precision (AP@30, AP@50 and AP@70 respectively) in relation to the volume of communication in log2. The baseline is drawn in a diamond shape, early fusion in a triangle, late fusion in square and our contributions in stars and in a plus shape. The results on V2XSet are based on fill markers and on empty markers for OPV2V. Configuration 1 (Config. 1) takes into account the data filtering module with a channel dimension of [8,32,64] instead of [64,128,256]. Configuration 2 (Config. 2) takes into account a channel dimension of [8,64,128]. Both configurations take into account the new message representation format to be exchanged between connected agents.

The results highlight that late fusion provides higher accuracy than the baseline for all average precision thresholds with a median communication volume of around 2^9 bytes (B). The AP is close to the same accuracy as the early fusion (EF) on both datasets. The communication volume for the EF configuration is much higher with around 2^{21}B. In AP@70, the accuracy of late fusion decreases by 10% compared to the early fusion approach but sill higher than the baseline by around 10% to 20% depending on the dataset.

Comparison of execution times in *ms* for each part of the proposed pipeline compared to the baseline [4]. The first two best results for each datasets are highlighted in bold.

Model	Dataset	Encoder	Confidence Generator	Message Exchange	Message Packing	Decoder
Where2comm [4]	V2XSet	4.82	**0.45**	31.05	0.12	**0.14**
	OPV2V	5.25	0.44	34.59	**0.12**	**0.15**
Config. 1	V2XSet	**4.74**	0.46	**15.39**	0.12	0.15
	OPV2V	**5.17**	0.46	**18.10**	0.13	0.16
Config. 2	V2XSet	5.25	**0.45**	19.26	**0.10**	0.15
	OPV2V	5.57	**0.43**	25.57	**0.12**	**0.15**

Our contribution yields results comparable to those of intermediate fusion, aiming to achieve a communication volume similar to late fusion while maintaining an average precision close to that of early fusion. The first configuration reduces the communication volume of the baseline by 91% on both datasets and the second configuration by 86% and 90% on V2XSet and OPV2V respectively. Those results are based on the $\log_2$ base conversion to bytes. In terms of accuracy, a lower channel dimension to extract the most relevant features as in the first configuration increases the AP@30 and AP@50 up to 7% reaching 78% and 73% on V2XSet respectively. The accuracy on the OPV2V dataset has not been impacted and is similar to the baseline. On the other hand, the accuracy with the second configuration decreases on OPV2V by 9%, 3% and 9% on AP@30, AP@50 and AP@70 respectively.

Both configurations brings a trade-off between volume of communication and accuracy of 3D object detection. It highly depends on the use case and the type of classes to predict. For smaller objects, the configuration one is preferable, whereas for bigger objects, such as trucks, the configuration two is the one to choose.

Table 1 provides the execution time (in milliseconds) of each part of the proposed pipeline with both configurations compared to the baseline Where2comm. While the data filtering in the encoder does not significantly impact the execution time of the encoder part, the message representation module reduces the latency of the message exchange part by 50%. The whole pipeline runs on NVIDIA GeForce RTX 2080 Ti GPUs with a latency of around 20.86ms compared to 36.58ms for the baseline, which represents a processing of 48 FPS compared to 27 FPS respectively.

Conclusion and perspectives

This work proposes new optimization strategies for the data exchange in the context of collaborative perception in a V2X manner. The developed pipeline, based on the Where2comm algorithm, reduces the bandwidth consumption while maintaining the performances of 3D object detection with two key components: the data reduction module and the representation of messages to be transmitted between agents.

Experimental results show that the proposed approach achieves a trade-off between accuracy and consumption based on LiDAR processing with a data reduction of 89.77% and 92.19% on V2XSet and OPV2V respectively. The accuracy of the perception performances have been increased by 5% on AP@50 compared to the baseline. Real-time performance for 3D object detection has been achieved using a NVIDIA GeForce RTX 2080 Ti GPU with a latency of 20.86ms for the whole pipeline, which reduces the baseline execution time by 15.72ms.

Future works will be focused on the development of a more robust strategy that optimizes bandwidth usage even with a large volume of data, while maintaining high perception performance. Indeed, our method shows a significant advantage in terms of bandwidth consumption when the amount of information to be transmitted is limited. However, as this number increases, so does the bandwidth consumption. It is also essential to conduct an in-depth study to compare the state-of-the-art recommended data fusion methods with those emerging from our experiments. Such an analysis would make it possible to identify the most efficient approaches in terms of bandwidth management according to the volume of information processed. The literature suggests that intermediate fusion is efficient for its accuracy and consumption trade-off. Our works indicate that late fusion is preferable on the simulated datasets used. Proposed collaborative approaches need to be tested on real-world datasets to assess their impact on a collaborative system performance.Therefore, there is a strong need for available real-world datasets that cover both V2I and V2V scenarios.

Acknowledgments

The authors acknowledge the infrastructure and the support of the SCARLET team working at the VEDECOM Institute.

References

[1] Siheng Chen, Baoan Liu, Chen Feng, Carlos Vallespi-Gonzalez, and Carl Wellington. 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. *IEEE Signal Processing Magazine*, 38(1):68–86, 2021.

[2] Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine*, 15(6):131–151, 2023.

[3] Yedid Hoshen. Vain: Attentional multi-agent predictive modeling. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[4] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4874–4886. Curran Associates, Inc., 2022.

[5] Yue Hu, Shaoheng Fang, Weidi Xie, and Siheng Chen. Aerial monocular 3d object detection. *IEEE Robotics and Automation Letters*, 8(4):1959–1966, 2023.

[6] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2023.

[7] Yue Hu, Xianghe Pang, Xiaoqi Qin, Yonina C. Eldar, Siheng Chen, Ping Zhang, and Wenjun Zhang. Pragmatic communication in multi-agent collaborative perception. *ArXiv*, abs/2401.12694, 2024.

[8] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12689–12697, 2019.

[9] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc.

[10] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4105–4114, 2020.

[11] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883, 2020.

[12] Eloi Mehr, Ariane Jourdan, Nicolas Thome, Matthieu Cord, and Vincent Guitteny. Disconet: Shapes learning on disconnected manifolds for 3d editing. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3473–3482, 2019.

[13] Sainbayar Sukhbaatar, arthur szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[14] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel 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*, page 605–621, Berlin, Heidelberg, 2020. Springer-Verlag.

[15] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[16] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi 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)*, page 2583–2589. IEEE Press, 2022.

[17] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 25151–25164. Curran Associates, Inc., 2023.

[18] Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. Spatio-temporal domain awareness for multi-agent collaborative perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23326–23335, 2023.

[19] Kun Yang, Dingkang Yang, Jingyu Zhang, Hanqi Wang, Peng Sun, and Liang Song. What2comm: Towards communication-efficient collaborative perception via feature decoupling. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 7686–7695, New York, NY, USA, 2023. Association for Computing Machinery.

[20] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *ArXiv*, abs/1904.07850, 2019.

Author Biography

Besma Abdali received the M.S. degree in mobile autonomous systems from the University of Paris-Saclay, France in 2024. Her studies has been focused on the control, command and perception of autonomous systems.

Quentin Picard received the Ph.D. degree in 2023 in the perception of mobile autonomous systems from the University of Paris-Saclay at the CEA LIST Institute, Saclay, France in collaboration with the IBISC (Computer Science, Bio-Informatics and Complex Systems) laboratory. He is currently a research engineer at the VEDECOM Institute, Versailles, France with the CCAM team for onboard and offboard perception. His research interests involve real-time processing for embedded systems and V2X cooperative perception.

Maryem Fadili received her engineering degree in land surveying from l'École Supérieure des Géomètres et Topographes du Mans, France in 2015. She has held various research positions focused on remote sensing and 3D mapping. Since 2018, she has been working as a research engineer at the VEDECOM Institute, where she began pursuing her PhD in signal and image processing in 2024. Her research interests include environment perception in autonomous driving and data fusion.