ENPH 455/555 Final Report

Cooperative Perception for Autonomous Vehicles

by

G. Clark

A thesis submitted to the

Department of Physics, Engineering Physics, and Astronomy

in conformity with the requirements for

the degree of Bachelor of Applied Science

Queen's University

Kingston, Ontario, Canada

April 2023

Abstract/Executive Summary

The demand for autonomous vehicles (AVs) has seen a significant rise in recent years, with the global AV market expected to reach $400 billion by 2030. Ensuring the safety and reliability of AVs is of paramount importance, and cooperative perception, which involves collaboration between multiple vehicles and roadside infrastructure sensors, has the potential to improve both. This study aims to investigate the impacts of different feature fusion methods on real-world V2X datasets, focusing on the accuracy and data expenditure of these models. The paper provides an in-depth analysis of cooperative perception systems, including feature encoding, feature extraction, feature fusion, and object detection. The DAIR-V2X dataset will be used to train and test cooperative perception models, and the results will be compared with other fusion methods. The objective is to provide a comprehensive understanding of the advantages and limitations of intermediate feature fusion in cooperative perception systems, thus contributing to the development of safer and more efficient AVs. The goal was to implement Spatial- and Channel-adaptive intermediate feature fusion into the provided DAIR-V2X codebase. Unfortunately, due to limitations in the base code, we switched to a recently published cooperative perception model called CoAlign, which offers intermediate collaboration for multiple datasets, including DAIR-V2X. CoAlign effectively addresses the pose errors inherent in intermediate feature fusion and implements a pose correction phase using agent-object pose graph modeling. Upon successful implementation of the multiple PointPillar based models created by CoAlign, we compared the performance of early, late, and intermediate feature fusion methods in terms of average precision (AP) and average byte (AB) scores. The results demonstrated the superiority of intermediate feature fusion for cooperative perception in autonomous vehicles (AVs). They also illustrated the vast outperformance of the CoAlign model as compared to the base DAIR model. Future work includes refining the DAIR-V2X codebase and investigating alternative fusion methods to advance the understanding and performance of cooperative perception models for AVs. Further exploration of various communication modalities may also contribute to the development of more robust and efficient cooperative perception systems.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Farhana Zulkernine and Donghao Qiao for their support and guidance throughout the year. Their extensive knowledge and proficiency in their respective domains have been invaluable. I would like to thank Aaron Visser who diligently oversaw the Queen’s Computer Cluster. I would also like to thank the rest of the BAMLAB researchers for their and help throughout the year.

1 Introduction and Motivation

In recent years, the drive to mainstream autonomous vehicles (AVs) has gained significant momentum in technology hubs across the globe. A 2019 compendium by McKinsey predicts that the global AV market will rise to an estimated $400 billion by 2030, a quarter of the total worldwide vehicle market [1]. Technology giants and automakers like Tesla, Alphabet's Waymo, and GM's Cruise have invested tens of billions of dollars into creating and releasing their own AVs for public and private use [2]. Yet even in the five states currently testing and deploying AVs, it is already clear that there is still much work to be done before widespread implementation [3] [4]. Ensuring the safety of passengers and pedestrians is a critical concern in the development of AVs. Mishaps such as the 2018 Arizona Uber accident and the 2016 Tesla Autopilot crash highlight the importance of developing reliable and accurate perception systems for these vehicles [5]. Cooperative perception is one approach that can help improve the safety and efficiency of AVs [6]. As the demand for AVs continues to grow, it is essential to continue researching and improving the technology to make self-driving cars safer for every stakeholder involved.

Cooperative perception is a concept in autonomous driving that involves the collaboration between multiple vehicles and roadside infrastructure sensors to share sensor data and create a more accurate understanding of the driving environment [7]. Cooperative perception is a particularly beneficial model as it solves many of the safety problems that commonly lead AVs to make mistakes. AVs rely heavily on perception systems to navigate through their environment. However, traditional perception systems such as cameras and LiDAR have many limitations, especially in adverse weather conditions or in areas with poor lighting [8] [9]. Even in perfect conditions, single vehicle sensors have limitations in range, and inherent error that create opportunities for mistakes. Cooperative perception can help overcome these challenges by utilizing data from multiple sources resulting in higher accuracy, increased spatial information, and increased redundancy [10]. This paper will investigate the impacts of different feature fusion methods on real world V2X datasets to compare the accuracy and data expenditure of these models.

2 Background

2.1. Autonomous Driving

According to the International Society of Automotive Engineers (SAE) there here are six classes of autonomous driving [11]. These distinctions are made through ordered levels of autonomy from 0, being no autonomy, to 5, being complete autonomy. A level five AV hands complete control over to the vehicle and has no need to provide a steering wheel, accelerator, or braking system for the passengers. Level 5 AVs will drive in any environmental and roadway conditions. AVs currently on the road rely on their own sensors to perceive the space around them [12]. This perpetuates the current downfalls outlined in Introduction and Motivation and prevents the widespread implementation of fully autonomized vehicles [13].

2.2. Vehicular Communication

Vehicular communication is the foundation of cooperative perception. For vehicles to obtain a comprehensive understanding of their space, they must take information from other sources and points of view. There are two main types of vehicular communication, Vehicle-to-Vehicle (V2V) and

Vehicle-to-Infrastructure (V2I). It is widely agreed that to implement Level 5 AVs, both types of vehicle communication are required [14]. This comprehensive communication is aptly named Vehicle-to-Everything (V2X). Vehicle to everything requires a far more involved feature fusion process due to discrepancies including sensor types, GPS localization noise, coordinate differences, etc. [8].

2.3. Cooperative Perception

The perception of a vehicle can be thought of as the ability for a vehicle to establish an understanding of its surroundings using sensors, cameras, and other relevant data. Cooperative perception takes the perception of each vehicle and/infrastructure in the vicinity and combines them for a comprehensive view of the space for each member sharing their data [15]. Cooperative perception uses point cloud data and takes place over four main steps: feature encoding, feature extraction & projection, feature fusion, and object detection.

2.3.1. Point Clouds

Point clouds are the basis of much of the data used for 3D object detection and will be the base of cooperative perception. 3D representations of real-world objects and environments created using a large set of discrete points in space. This kind of data is the primary source used for most object detection models for AV's due to the three-dimensional data that is retrieved from the LiDAR sensors. Three additional dimensions are stored for feature projection. These attributes describe the orientation and include roll, yaw, and pitch. These combined position and orientation information is called pose. Examples of point clouds visualized can be seen in Figures Figure 3,Figure 5, and Figure 6.

2.3.2. Feature Encoding

To extract features from the LiDAR sensors, a pseudo-image must first be made from the point cloud. While there are many ways to get from A to B, one of the fastest and most effective encoders come in the form of PointPillars [16, 17]. These benefits present a clear leader in feature encoding: the Pillar Feature Net. A single point in a point cloud has 4 attributed coordinates: *x, y, z, r.* PointPillars are arbitrarily formed based on a 3D grid over the point cloud. Once the pillars are created, each point gains 5 dimensional coordinates: x_c, y_c, z_c, x_p, y_p where the c subscript represents the distances to the arithmetic means of each point in the pillar and the p subscript represents the distance from the pillar x, y center. The sparsity of points in these pillars is leveraged to create the tensor (D x P x N) where D is the dimension of the data, P is the number of non-empty pillars per sample and N is the number of points per pillar. Further convolutions provide us with a (C x P x N) tensor and are followed by max pooling resulting in a (C x P) tensor. Pillars with C features are projected onto their original locations to create a pseudo-image which will be down sampled to create the tensor (C x H x W) where H and W are the height and width of the projected image [18] [19]. This process can be seen in Figure 1.

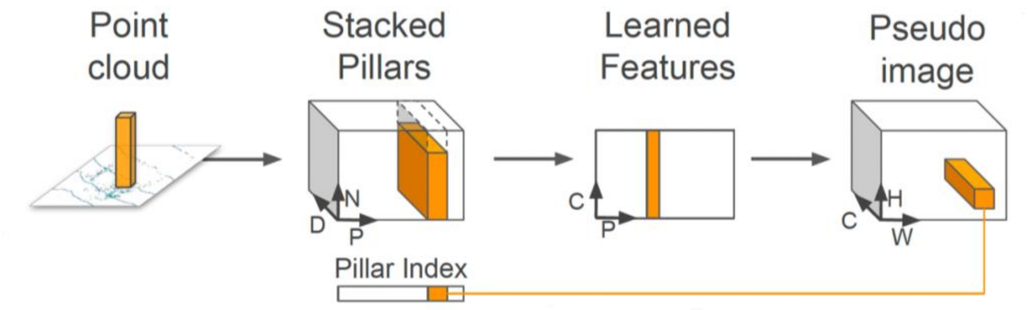

Figure 1 visualizes the PointPillar feature encoding process [16].

2.3.3. Feature Extraction & Projection

A Feature Pyramid Network [20] [10] is used in feature extraction to systematically extract features using two counteracting routes. The top-down route produces features at increasingly lower resolution using 2D convolution, batch normalization and ReLU. Each of the three low resolution maps are then up sampled and concatenated resulting in a feature map of the features contained in the original point cloud.

As the vehicles and infrastructure have different POVs, the coordinate systems of the surrounding vehicles/infrastructure must align with the coordinate system of the receiving vehicle/infrastructure. This projection and alignment of coordinates is possible due to the LiDAR pose information – 3D spatial and orientation coordinates.

2.3.4. Feature Fusion

Feature fusion is a process added to a network for more precise object detection within multiple sets of images from different points of view [21]. There are three different methods for feature fusion currently in use in common literature. Early/low-level fusion fuses together raw data without any of the preprocessing done in 2.3.3. This method, while accurate, requires massive amounts of data to be transferred between vehicles and infrastructure [5]. Late fusion waits to fuse features until they are detected by the detection head [22]. This model requires a much less intensive data transfer yet tends to lose the accuracy that comes with more information shared before detection. Intermediate/feature-level fusion mitigates the size of the necessary data transfer while maintaining high accuracy and complexity in fusion [19] [20] [9]. Intermediate fusion takes the extracted features and maps them to 4D tensors to be further concatenated. The architecture of the cooperative perception using intermediate feature fusion is visualized in Figure 2. This fused map is them passed to the subsequent layers of the model for object detection.

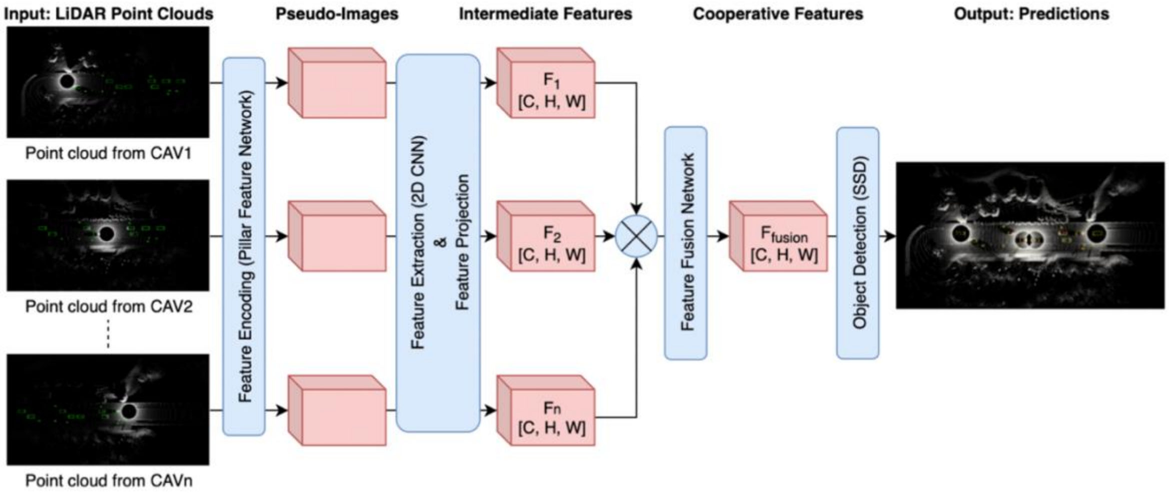

Figure 2 displays the architecture of cooperative perception using intermediate feature fusion [19].

2.3.5. Object Detection

Object detection is the final step in the cooperative perception process. With the fused feature representation, the actual object detection process takes place. Various machine learning and deep learning models, such as SECOND [23], PointPillars [16] [18], etc. can be used to identify and localize objects in the environment [24]. These fused tensors are sent through a Single Shot Detector where the predictions are compared to ground truth boxes. There are multiple objects that need to and can be detected by many cooperative perception models including but not limited to: cars, trucks, bicycles, pedestrians, etc. [10].

3 Problem Definition

The objective of this paper is to investigate the potential benefits and trade-offs of applying cooperative perception using intermediate feature fusion to the DAIR-V2X dataset. This study aims to implement the intermediate feature fusion model detailed in [19] into the DAIR-V2X codebase. The results of the intermediate fusion model will then be compared to early and late fusion models presented in [10], focusing on their performance in terms of accuracy and memory usage. To achieve this, two feature extraction models will be employed: Imvoxelnet for image data and PointPillar for point cloud data. By comparing the results with other fusion methods applied to DAIR-V2X, we aim to provide a comprehensive understanding of the advantages and limitations of intermediate feature fusion in cooperative perception systems.

The DAIR-V2X dataset will be used to train and test the model which will be implemented using the PyTorch framework [10]. This dataset comprises 70,000 point clouds from lidar sensors and camera frames. Examples of this data can be seen in Figure 3. This dataset was chosen over other more popular datasets as it is the only publicly available dataset that is both real world and supports vehicle to everything. Bounding boxes on this dataset span a range of different types of objects and includes cars, trucks, vans, busses, pedestrians, cyclists, tricyclist, motorcyclists, and traffic cones.

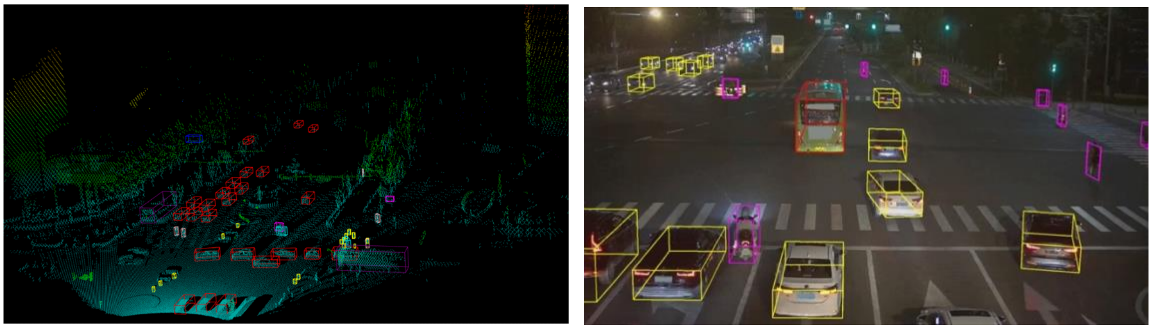

Figure 3 shows one LiDAR frame and one image frame from the extensive DAIR-V2X dataset. These images have the labeled bounding boxes around them and are both from the infrastructure view.

Cooperative perception models are generally evaluated on their Average Precision (AP). AP is calculated at different intersection-over-union (IoU) thresholds (between 0.1 and 1) this dataset will be evaluated at 0.5 and 0.7 In object detection, labeled data will have a box called the ground truth box to indicate an object to be detected. The object detection model will then predict a box of its own. IoU is calculated by dividing the area of intersection of the boxes by the total area of the boxes. Similar models have achieved up to 64% AP@0.5. It is expected that the model will obtain a value close to or above 64% due to the implementation of intermediate fusion. Average byte (AB) is a common metric for measuring the memory expense of certain computations. This measurement represents the average number of bytes (8 bits of data) transferred in the form of raw data, intermediate data representation, object level outputs, and other miscellaneous information [25]. Published AB values span currently span the range of 306.79 to 1.36M for late fusion and early fusion respectively. It is expected that the value will be somewhere on the lower end of this range. As the type of feature fusion has larger impact on the memory expense than time expense, the time scale will not be an evaluation metric [26].

4 Methods

4.1. DAIR-V2X Model

DAIR-V2X is an extensive cooperative perception model based on 3D object detection through LiDAR point clouds images. The model uses ImvoxelNet for 3D object detection from images. ImvoxelNet focuses on leveraging monocular images, which are more cost-effective and widely available. The model achieves this by projecting 2D image features into a 3D voxel space while preserving the semantic information present in the image [27]. The PointPillar (Feature Encoding) and SECOND [23] models are used and for 3D object detection from point clouds. SECOND leveraging a sparse convolutional neural network (CNN) for processing the irregularly structured point cloud data. SECOND is designed to work efficiently on large-scale point cloud datasets such as DAIR-V2X. The DAIR cooperative perception model supports both early and late feature fusion, though doesn't offer intermediate. The codebase uses mmDetection3D for training and evaluation.

4.2. Implementation

4.2.1. DAIR-V2X

The python written DAIR-V2X open-source codebase is an intensive model. Implementation is a longer process as there are many system configurations to go through before training and testing. As forementioned, mmDetection3D is used to train and test the models as DAIR has not yet implemented this themselves. Pypcd is another library that is necessary for the process as it manages the pointcloud aspect of the model. There is a range of different model configurations that can be chosen from for training and testing. The available fusion methods include vehicle-only, infrastructure-only, early, and late. ImvoxelNet is used for images and PointPillars for point clouds.

Training and testing the models starts with converting the dataset into a new format. The KITTI dataset is a benchmark in the AV and computer vision field and has been referenced in thousands of articles since its release in 2012 [28] [29]. Due to its prominence, the format of testing and training has been fit to its dataset and therefor the DAIR format must be converted. Once the conversion is done, scripts using the proper config files corresponding to training and evaluation are run to assure the model is giving similar results to those published.

4.2.2. Intermediate Fusion

The code and background for implementing the intermediate fusions is based on the work achieved in [19]. Spatial-wise and Channel-wise Adaptive feature Fusion (S, C-AdaFusion) are the two types of fusion methods used for intermediate fusion in this paper.

S-AdaFusion uses spatial features to fuse feature maps to be sent to the object detection path. Spatial features are generated by max or average pooling applied to the input feature map of dimensions $F \in R^{n \times C \times H \times W}$. These operations condense the map into separate $S_{avg}, S_{max} \in R^{1 \times C \times H \times W}$ tensors. These separate maps are concatenated together, forming a 4D tensor of form $F_{spatial} \in R^{2 \times C \times H \times W}$. Finally, volumetric segmentation is done through a 3D convolution followed by a ReLU activation function. This step brings the final fused feature map in the form of $F_{fusion} \in R^{1 \times C \times H \times W}$.

C-AdaFusion leverages the channels on the 4D input tensor F and uses a 3D CNN to extract channel features. The number of input channels on the CNN should equate to the number of CAV/Infrastructure sensors which combine feature maps. In the case of this model, a CAV and an infrastructure sensor will be used for fusion. Like S-AdaFusion, C-AdaFusion applies average and max pooling to extract channel descriptors which describe the different aspects of the input data. These become $C_{max,avg} = R^{n \times 1 \times 1 \times 1}$ and get concatenated before passing through two dense layers with linear ReLU and Sigmoid activation functions. Channel reduction is then applied, yielding a final fused feature map of $F_{fusion} \in R^{1 \times C \times H \times W}$.

Implementing either of these intermediate fusion methods will be done through creation and editing of configuration files in the codebase to properly transform and fuse the data in the correct way at the correct time.

5 Iteration and Results

5.1. DAIR-V2X

During the process of implementing intermediate fusion in the DAIR-V2X codebase, several challenges and issues were encountered. While the DAIR codebase provided a comprehensive tutorial on the execution of their code, many of the errors encountered, once solved, would yield separate errors. Sustained attempts were made to understand and solve these errors including reaching out to each author individually and posting multiple issues on the public GitHub forum. Despite these diligent efforts, it became increasingly clear that the base code itself was fraught with difficulties and limitations that hindered smooth integration and execution. This realization led to an assessment of the situation and ultimately, the decision to move on from the current approach.

5.2. CoAlign

CoAlign, a cooperative perception model applicable to multiple databases including DAIR-V2X, was released in March of 2023. The CoAlign model highly resembles the cooperative perception model outlined in Background [30]. The two main differences come in the focus on accurately detecting the position and orientation (pose) of the objects and the availability for intermediate collaboration for each of the supported datasets. This model addresses the inherent pose errors that come with the use of intermediate feature fusion and implements a pose correction phase. This is done locally through an agent-object pose graph, each node is associated with a pose, shown in Figure 4.

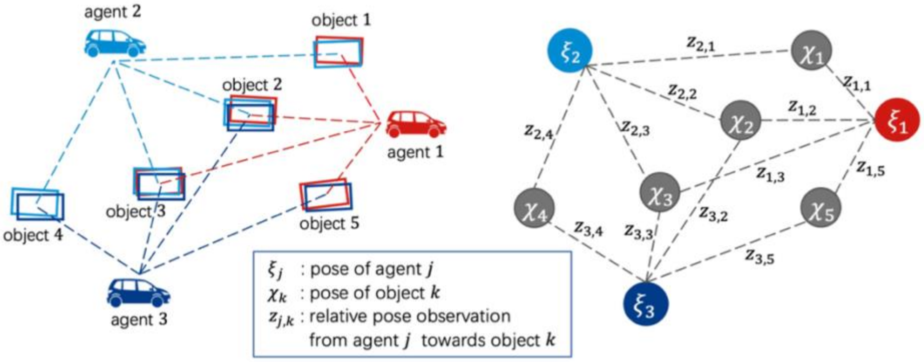

Figure 4 shows an example of an agent-object pose graph [30].

In this agent-object pose graph, each node is associated with a pose. Ideally, the pose of an object should be consistent from the views of multiple agents. To promote this pose consistency, the Simultaneous Localization and Mapping (SLAM) optimization problem. SLAM is a robotics technique that concurrently constructs a map of an unknown environment and estimates the agent's position within it, using data from multiple sensors for real-time navigation and exploration. However, the difference lies in the fact that graph-based SLAM aligns the same object across multiple time stamps, while this model aligns the same object at the same time stamps yet detected by multiple agents. This optimization problem is then solved using the Gaussian-Newton or Levenberg-Marquardt algorithms. With the corrected relative pose, the feature maps from different agents can be aligned, which allows the fusion of features at an intermediate level.

Further optimization of the DAIR-V2X dataset includes the creation of additional 3D box annotations to enable a full 360^o detection. After successfully implementing the multiple PointPillar based models created by CoAlign, the AP results are compared to the published DAIR-V2X cooperative perception model in the Table 1. These results clearly show that the intermediate feature fusion models created using CoAlign show a far superior AP performance over early or late fusion while maintaining an AB score comparable to late fusion.

Table 1 shows the published results of the DAIR-V2X paper compared to the experimental results obtained from the CoAlign code base.

Model Design	***Fusion Method***	***Model Type***	***Dataset***	***AP (0.5 IoU)***	***AP (0.7 IoU)***	***AB (bytes)***
DAIR-V2X [10]	Early	PointPillars	DAIR-V2X Sync	53.73	N/A	1382275.75
	Late		DAIR-V2X Sync	47.96	N/A	336.16
CoAlign [30]	Early		DAIR-V2X 360°	0.729	0.596	~ 1000
	Late		DAIR-V2X 360°	0.716	0.573	~ 1000000
	Intermediate		DAIR-V2X 360°	**0.772**	**0.627**	~ 10000

The inaccuracy in the AB Coalign results comes from the general calculations implemented compared to the DAIR-V2X specific byte calculations. A visualization of the AB results of a partially and fully trained model are shown in Figure 5 and Figure 6. The red boxes are the model predictions whereas the green boxes are the labels. The small numbers next to the red predicted boxes indicate the IoU score.

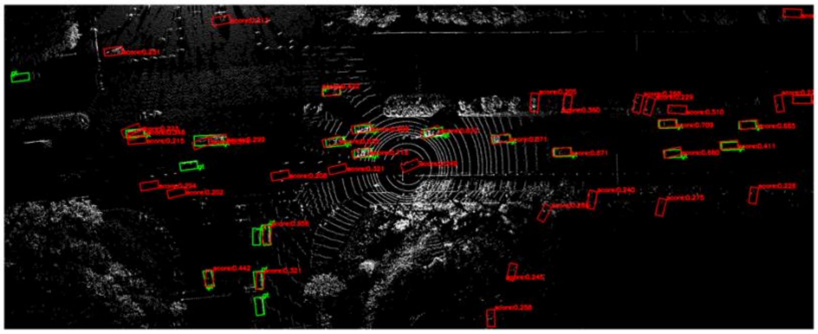

Figure 5 shows the results of a partially trained intermediate fusion model (5 epochs).

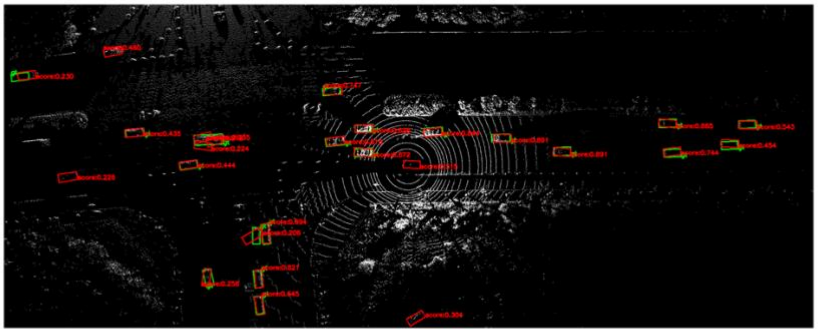

Figure 6 shows the results a fully trained intermediate fusion model (30 epochs).

6 Discussion

6.1. Results

The results shown in Table 1 clearly show the superiority of intermediate fusion in average precision. While there are no published AP@0.7 statistics for the DAIR early and late fusion model, each CoAlign fusion method vastly outperforms the former even in AP@0.7 results. The outperformance of non-intermediate fusion models is a result of the novel agent-object pose graph modeling. This results in enhanced pose consistency between collaborators in each of the fusion methods. The AB results for the CoAlign values are general, the sample calculation provided by the codebase is accurate to the thousandth byte. These results are still in line with the prediction of intermediate fusion being on the lower end of the spectrum between early and late fusion.

6.2. Economic Analysis

This project has a low economic cost due to the open-source robust real-world dataset and a computing cluster for data storage and model running provided by the Queen's Computer Cluster. Implementing cooperative perception for Level 5 AVs involves significant infrastructure expenses and ongoing upkeep [31]. Benefits include increased traffic throughput, reduced energy consumption [32], decreased traffic accidents, insurance costs, and vehicle maintenance expenditures [13]. However, AV adoption also presents challenges as an infrastructure project of this magnitude will cost hundreds of billions of dollars [33] [34]. The technology's widespread deployment will profoundly impact urban planning, infrastructure investments, and overall transportation costs, creating a safer, more efficient transportation ecosystem.

6.3. Environmental Considerations

The combination of electric vehicles and AVs could lead to a green revolution in the automotive industry. Private passenger vehicles contribute over 60% of transportation sector greenhouse emissions. By optimizing routes, speeds, and traffic flow, AVs can decrease emissions significantly [35]. However, the energy required for data transfer and analysis, as well as production, operation, and disposal of these vehicles, involves substantial energy consumption and natural resource use, particularly for battery creation [33].

6.4. Ethical Considerations

Level 5 AV adoption presents ethical implications, including significant job loss in the US transportation industry, which constitutes 3.5% of total GDP and pays around $500 billion in worker compensation annually [33]. However, AVs could reduce traffic accidents by up to 90%, potentially saving thousands of lives each year [13] [36]. Balancing mass unemployment risks with passenger safety improvements presents an ethical challenge. Additionally, vast data collection by AVs poses privacy and security risks, necessitating robust privacy protection measures and ethical guidelines to maintain public trust in AV technologies.

6.5. Equity Considerations

Equitable access to transportation is a critical aspect of social justice, and the introduction of level 5 AVs with cooperative perception could both create and mitigate disparities in transportation services. AVs have the potential to improve transportation access for those with disabilities, the elderly, and those without access to personal vehicles by providing efficient and affordable transportation alternatives. On the other hand, the initial costs of AV technology and infrastructure investments could lead to unequal distribution of benefits, with poor populations potentially being left behind or disproportionately burdened with negative externalities. The outcomes of these issues will largely be decided by the policy makers and their ability to create an equitable landscape ahead of the AV revolution [37].

6.6. Risks and Safety

Cooperative perception improves autonomous vehicle (AV) performance but introduces safety risks [38]. AVs rely on real-time data exchange, where data integrity, accuracy, and timeliness are crucial. Risks involve data tampering, hacking, or transmission errors, potentially leading to incorrect decisions and safety threats [39]. Communication latency and network congestion can cause outdated or delayed information, increasing collision risks. Addressing these risks requires robust security protocols, fault-tolerant systems, and reliable communication infrastructures [37].

7 Limitations

This project encountered several limitations that may have impacted the overall progress and results. Firstly, the implementation of feature fusion methods was not fully achieved, primarily due to issues with the code. Despite dedicating a significant amount of time attempting to debug and resolve these issues, both pertaining to the system and the published code itself, the model remained problematic and hindered the proper execution of feature fusion techniques. This limitation restricted the ability to thoroughly explore and analyze the effects of different fusion strategies on the model's performance. Secondly, while the Queen's server generally provided satisfactory performance, there were instances when access to GPUs was limited, or the server experienced downtime. These interruptions in server availability resulted in delays in the research process and may have affected the ability to optimize the model and refine the experimental setup. Lastly, the accessibility and size of the database posed additional challenges. The database, consisting of over 60 GB of data, is only accessible from mainland China, which created hurdles in obtaining the necessary data for the study. The large size of the database further exacerbated the issue, as it complicated the data acquisition process and increased the time required for data transfer and storage.

8 Conclusion

This project set out to implement an intermediate feature fusion network into the DAIR-V2X codebase. The comparison between early, late, and intermediate fusion by means of AP and AB scores would point to the strongest fusion method for cooperative perception. Unfortunately, due to unforeseen setbacks with the DAIR-V2X codebase, the intermediate feature fusion model outlined in Intermediate Fusion was unable to be implemented. Luckily, a recently published paper was able to complete the task of applying intermediate cooperation to the DAIR-V2X dataset. The CoAlign codebase was a far better code both in ease of implementation and ease of implementation and in quality of results. The CoAlign model vastly outperformed (in terms of AP) DAIR's model even when comparing early and late feature fusion. The AB results generally align with predictions, but a more accurate calculation would need to be done to complete a full comparison. If given more time, getting better AB results would be the first task. More generally, working with the DIAR-V2X code to implement intermediate feature fusion method outlined in Intermediate Fusion would continue to be the main goal.

Future work could explore the effects of various communication modalities, such as cellular networks or dedicated short-range communication (DSRC), on the performance of cooperative perception models. Additionally, testing the models under different driving scenarios or environmental conditions could provide valuable insights into their robustness and adaptability. Other important future work ties into the privacy and ethics of cooperative perception in the form of exploring data privacy and security measures. Future research could focus on the development of secure communication protocols, robust encryption methods, and data anonymization techniques to mitigate potential risks and vulnerabilities.

For any further advancements in this area of research, it is important to continue considering the environmental, ethical, equitable implications as well as maximizing safety for each stakeholder.

References

[1] McKinsey Center for Future Mobility, "The future of mobility is at our doorstep," 2019/2020. [Online]. Available: https://www.mckinsey.com/~/media/McKinsey/Industries/Automotive%20and%20Assembly/Our%20Insights/The%20future%20of%20mobility%20is%20at%20our%20doorstep/The-future-of-mobility-is-at-our-doorstep.ashx.

[2] R. Fannin, "Where the billions spent on autonomous vehicles by U.S. and Chinese giants is heading," CNBC, 21 May 2022. [Online]. Available: https://www.cnbc.com/2022/05/21/why-the-first-autonomous-vehicles-winners-wont-be-in-your-driveway.html. [Accessed 17 February 2023].

[3] L. Wood, "Global Autonomous Vehicles Market Report 2022: Tech Promises to Increase Road Safety and Driver Comfort - ResearchAndMarkets.com," Business Wire, 13 October 2022. [Online]. Available: https://www.businesswire.com/news/home/20221013005688/en/Global-Autonomous-Vehicles-Market-Report-2022-Tech-Promises-to-Increase-Road-Safety-and-Driver-Comfort---ResearchAndMarkets.com. [Accessed 27 February 2023].

[4] H. G. Seif and X. Hu, "Autonomous Driving in the iCity—HD Maps as a Key Challenge of the Automotive Industry," *Engineering,* vol. 2, no. 2, 2016.

[5] Q. Chen, S. Tang, Q. Yang and F. Song, "Cooper: Cooperative Perception for Connected Autonomous Vehicles based on 3D Point Clouds," in *39th IEEE International Conference on Distributed Computing Systems*, Dallas, 2019.

[6] S.-W. Kim, B. Qin, Z. J. Chong, X. Shen, W. Liu, E. Frazzoli and D. Rus, "Multivehicle Cooperative Driving Using Cooperative Perception: Design and Experimental Validation," *IEEE Transactions on Intelligent Transportation Systems,* vol. 16, no. 2, pp. 663 - 680, 2015.

[7] P. Lv, J. Han, Y. He, J. Xu and T. Li, "Object Perceptibility Prediction for Transmission Load Reduction in Vehicle-Infrastructure Cooperative Perception," *Sensors,* vol. 22, no. 11, 2022.

[8] R. e. a. Xu, "V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer," arXiv, Los Angeles, 2022.

[9] 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," Arxiv, 2022.

[10] 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.

[11] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," SAE Mobilus, 2016.

[12] E. Tanenblatt, M. Malterer and P. Stockburger, "Global Guide to Autonomous Vehicles," Dentons, 2023.

[13] J. Wang, H. Huang, K. Li and J. Li, "Towards the Unified Principles for Level 5 Autonomous Vehicles," *Engineering,* vol. 7, no. 9, pp. 1313-1325, 2021.

[14] X. Gan, H. Shi, S. Yang, Y. Xiao and L. Sun, "MANet: End-to-End Learning for Point Cloud Based on Robust Pointpillar and Multiattention," *Wireless Communications and Mobile Computing,* 2022.

[15] J. Guo, "Machine-Learning-Enabled Cooperative Perception on Connected Autonomous Vehicles," University of North Texas, Denton, 2021.

[16] A. H. Lang, S. Vora, H. Cesar, L. Zhou, J. Yang and O. Beijbom, "PointPillars: Fast Encoders for Object Detection from Point Clouds," 2018.

[17] J. Stanisz, K. Lis and M. Gorgon, "Implementation of the PointPillars Network for 3D Object Detection in Reprogrammable Heterogeneous Devices Using FINN," *Journal of Signal Processing Systems,* vol. 94, p. 659–674, 2022.

[18] J. JianWang, X. Guo, H. Wang, P. Jiang, T. Chen and Z. Sun, "Pillar-Based Cooperative Perception from Point Clouds for 6G-Enabled Cooperative Autonomous Vehicles," Huawei Technologies, Changchun, 2022.

[19] D. Qiao and F. Zulkerine, "Adaptive Feature Fusion for Cooperative Perception using LiDAR Point Clouds," 2022.

[20] Q. Chen, X. Ma, S. T. Tang, J. Guo, Q. Yang and S. Fu, "F-cooper: feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *4th ACM/IEEE Symposium on Edge Computing,* New York City, 2019.

[21] Y. Guo, Y. Xu and S. Li, "Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network," *Automation in Construction,* vol. 112, 2020.

[22] B. Hurl, R. Cohen, K. Czarnecki and S. Waslander, "TruPercept: Trust Modelling for Autonomous Vehicle Cooperative Perception from Synthetic Data," arXiv, 2019.

[23] Y. Yan, Y. Mao and B. Li, "Second: Sparsely embedded convolutional detection.," *Sensors,* vol. 18, no. 10, 2018.

[24] E. Arnold, M. Dianati, R. De Temple and S. Fallah, "Cooperative Perception for 3D Object Detection in Driving Scenarios Using Infrastructure Sensors," *IEEE,* vol. 23, no. 3, pp. 1852 - 1864, 2022.

[25] V.-I. C. 3. D. v. F. F. Prediction, "Yu, Haibao; Tang, Yingjuan; Mao, Jilei; Xie, Enze; Yuan, Jirui; Luo, Ping; Nie, Zaiqing," in *International Conference on Learning Representations,* Kigali, 2023.

[26] J. Gao, P. Li, Z. Chen and J. Zhang, "A Survey on Deep Learning for Multimodal Data Fusion," *Neural Computation,* vol. 32, no. 5, pp. 829-864, 2019.

[27] D. Rukhovich, A. Vorontsova and A. Konushin, "ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection," ARXIV, 2021.

[28] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[29] "KITTI," Papers with Code, [Online]. Available: https://paperswithcode.com/dataset/kitti. [Accessed 26 03 2023].

[30] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, C. Siheng and Y. Wang, "Robust Collaborative 3D Object Detection in Presence of Pose Errors," ARXIV, 2023.

[31] T. Litman, "Autonomous Vehicle Implementation Predictions," Victoria Transport Policy Institute, Victoria, 2023.

[32] G. Cui, W. Zhang, Y. Xiao, L. Yao and Z. Fang, "Cooperative Perception Technology of Autonomous Driving in the Internet of Vehicles Environment: A Review," *Sensors,* vol. 22, no. 15, 2022.

[33] L. M. Clements and K. M. Kockelman, "Economic Effects of Automated Vehicles," *Journal of the Transportation Research Board,* vol. 2606, no. 1, 2017.

[34] O. Baron, O. Berman and M. Nourinejad, "The Economics of Autonomous Vehicles," Rotman Business School, Toronto, 2019.

[35] Ó. Silva, R. Cordera, E. González-González and S. Nogués, "Environmental impacts of autonomous vehicles: A review of the scientific literature," *Science of The Total Environment,* vol. 830, 2022.

[36] Y.-K. Ou, Y.-C. Liu and F.-Y. Shih, "Risk prediction model for drivers’ in-vehicle activities – Application of task analysis and back-propagation neural network," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 18, pp. 89-93, 2013.

[37] K. Emory, F. Douma and J. Cao, "Autonomous vehicle policies with equity implications: Patterns and gaps," *Transportation Research Interdisciplinary Perspectives,* vol. 13, 2022.

[38] K. Othman, "Exploring the implications of autonomous vehicles: a comprehensive review," *Innovative Infrastructure Solutions,* vol. 7, 2022.

[39] S.-W. Kim and S.-W. Seo, "Cooperative Unmanned Autonomous Vehicle Control for Spatially Secure Group Communications," *IEEE Journal on Selected Areas in Communications,* vol. 30, no. 5, pp. 870 - 882, 2012.

[40] Q. Yang, S. Fu, H. Wang and H. Fang, "Machine-Learning-Enabled Cooperative Perception for Connected Autonomous Vehicles: Challenges and Opportunities.," *IEEE Network,* vol. 35, no. 3, pp. 96 - 101, 2021.