Global Research journal of Natural Science and Technology (Grjnst)
Volume: 03 - Issue 3 (2025)
p.ISSN: 2790-7643 e.ISSN: 2790-7651
www.grjnst.com

Enhancing Object Detection Accuracy in Occluded Scenarios Using V2X Cooperative Perception and Deep Learning

Ameer Hamza Nawaz

COMSATS University Islamabad, Attock Campus

nawazhamza464@gmail.com

Maria Soomro

MS Computer Science, Fast Nuces University, Karachi

mariasmro07@gmail.com

Nasir Ghaffar

PhD Scholar, Department of Mathematics, University of Central Punjab, Lahore

nasirghaffarmphilmath@gmail.com

Muhammad Rizwan Tahir

Department of Artificial Intelligence, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan.

therizwantahir@gmail.com

Abstract

Occlusion poses a significant challenge to accurate object detection in autonomous driving systems, particularly in dense urban environments where single-agent perception often fails. This research introduces a novel deep learning-based framework, **V2X-OccluFusion**, designed to enhance detection accuracy by leveraging Vehicle-to-Everything (V2X) cooperative perception and occlusion-aware pre training. The model combines self-supervised masked BEV feature reconstruction with lightweight state-space fusion architecture, enabling multi-agent vehicles to share and reconstruct spatially occluded information. Extensive experiments were conducted across diverse datasets and occlusion levels, including full visibility, partial, and heavy occlusion. Results show that V2X-OccluFusion significantly outperforms baseline models such as Early Fusion and V2X-ViT in both detection accuracy and object recall, especially under heavy occlusion, where it achieved a 25.5% performance improvement over baselines. Additionally, the model demonstrated lower GPU memory usage and faster inference speed, supporting real-time deployment. Communication efficiency was also superior, using less bandwidth while maintaining detection robustness under variable V2X conditions. These findings validate the effectiveness of combining cooperative multi-agent perception with occlusion-aware training for autonomous systems. The research contributes not only to improving detection under occlusion but also sets a foundation for scalable, real-time V2X perception systems adaptable to real-world constraints. The study concludes with recommendations for future enhancements involving multimodal data fusion and federated deployment.

Keywords*:* Autonomous driving, Cooperative perception, Deep learning, Object detection, Occlusion handling, V2X communication

Introduction

AVs continuously experienced challenges in object detection practice when occlusion happened, especially when the pedestrians or vehicles mounted behind the infrastructure or other road users. The response of researchers to this is the investigation of Vehicle-to-Everything (V2X) cooperative perception, where the many agents (comprising automobiles, infrastructure sensors) worked together to exchange sensor information in order to conquer obstructed fields of vision (Zhao et al., 2024; Li et al., 2024). Past work on BEV- or LiDAR-based agent-to-agent feature sharing had been shown to improve detection accuracy, but these approaches were not easily scalable, needed high bandwidth, and have unrealistic synchronization constraints.

At the same time, deep learning algorithms were used to disclose unavailable object information in covered areas, like masked feature reconstruction, and self-supervised pre-training methods (Zhao et al., 2024; Chen, 2025). The techniques demonstrated hopeful resistance against occlusion, although they were mostly tried in a clinically isolated location (single-vehicle environment). Thus it was critical that cooperative perception be integrated with occlusion-conscious deep learning in order to improve the robustness that is needed in realistic and occlusion-intensive, urban environments.

Other new systems such as CoMamba (Li et al., 2024) that scaled through state-space-model fusion to cooperatively fused cooperative perception systems and CooPre (Zhao et al., 2024) which uses a pretext proxy task with masking that gives occlusion robustness. At the same time, at the beginning of 2025, new cooperation forms of perception such as mmCooper, RGAtn, and V2XDGPE were introduced, which worked out gaps in areas, errors with the

pose, and economical movement with communications (Chen, 2025). Such developments highlighted the importance of blending self Hillikes point occlusion management and effective V2X blending in a single pipeline.

Research Background

Huang, et al. (2023) and Chen (2025) contextualized the importance of V2X systems as ways of improving situation awareness by determining the safety level of a vehicle based on the capabilities of the infrastructure and the vehicle itself. Cooperative schemes were divided into early, intermediate and late fusion strategies that had tradeoffs between latency, bandwidth and synchronization complexity. CoMamba was proposed by Li et al. (2024) and capitalized on bidirectional state-space models but more scalability when dealing with multi-agent information than transformer-based architecture-based networks. CoMamba was able to perform inference in real-time (~26.9 20 9 FPS), require only small amounts of memory (~0.64 24 9 9 GB), and its computational complexity could be scaled linearly with the number of agents. Comparing to standard benchmarks, CoMamba demonstrated outstanding performance on OPV2V and V2V4Real such that mean ~91.9 and 83.3 AP was obtained on IoU=0.5 and 0.7, respectively, being a lot better than the attention-based models (Li et al., 2024; TheMoonlight.io, 2024).

The proposed CooPre framework is a self-supervised pretraining model that is cooperative and BEV-guided in the aspects of masking and reconstruction between agents (Zhao et al., 2024). it conditioned 3D encoder to recreate missing data on LiDAR, encourages occlusion-robustness as well as enhance cross-domain generalization, and decreases dependency on ground-truth information. V2XV2V and OPV2V experiments of subsequent datasets based

on V2V4Real and V2XV2X datasets demonstrated stable performance improvements according to the variants of the scenario (Zhao et al., 2024).

By the end of 2024 and the beginning of 2025, new cooperative perception architecture, e.g., mmCooper and RGHeAttn and V2X-DGPE, was introduced, which further improved communication performance density, domain transferability and spatial-temporal stability (Chen, 2025). With such contributions, the issue of pose misalignment and scalability in urban V2X deployments was addressed. Furthermore, it was found that systems such as the infrastructure-enhanced perception system used in Singapore (~130 ms latency) could be deployed at the roadside to proactively detect occluded intersections in a practical way (Kim & Yu, 2025).

Research Problem

Although prior work had demonstrated the potential of cooperative perception to mitigate occlusion, existing V2X fusion frameworks faced trade-offs between performance, bandwidth, and real-time feasibility. Transformer-based fusion was powerful but computationally expensive and impractical with many agents; late-fusion or raw-model sharing reduced bandwidth costs but suffered accuracy losses (Li et al., 2024). Moreover, occlusion-specific occlusion-aware deep learning techniques and self-supervised pretraining had been developed mostly in isolation from cooperative perception contexts (Zhao et al., 2024). Thus, a gap remained in integrating scalable V2X fusion (with efficient state-space modeling) and occlusion-aware self-supervised feature reconstruction into a unified detection pipeline that delivered robust

performance under occluded urban conditions, while keeping bandwidth and latency within practical limits.

Research Objectives

1. To design and implement an intermediate-fusion V2X perception architecture using state-space modeling (CoMamba-style) that scaled efficiently with multiple agents.
2. To integrate masked feature reconstruction and BEV-guided self-supervised pre-training (inspired by CooPre) into the perception backbone to improve detection under occluded conditions.
3. To evaluate the unified framework using datasets such as OPV2V, V2V4Real, mmCooper, and real-world infrastructure-enhanced data, measuring 3D detection accuracy, occlusion-specific recall, latency, and communication usage

Research Questions

Q1. To what extent did self-supervised occlusion-aware pre-training increase 3D detection accuracy in occluded regions compared to cooperative baselines without pre-training?

Q2. Did the state-space-model-based fusion architecture maintain real-time performance and resource efficiency as the number of collaborating agents scaled up?

Significance of the Study

This study contributed a unified methodology that combined scalable V2X fusion with occlusion-resilient deep learning techniques, addressing a real-world obstacle in autonomous driving perception. The novel fusion-pretraining integration enhanced detection accuracy in occluded scenarios while preserving operational efficiency. Results potentially influenced the design of future large-scale cooperative perception systems, supporting safer autonomous vehicle deployment in complex urban environments.

Literature Review

V2X Cooperative Perception Architectures

The abovementioned problems and possibilities of V2X cooperative perception systems via architecture were described in several recent reviews. Huang et al. (2024) offered a categorization of early, intermediate and latefusion techniques are focused on agent heterogeneity, alignment, synchronization, communication constraints. Their point was that the conditions in V2X network were not always as ideal as it is expected, and the effectiveness of fusion was frequently impaired by the packet loss, delays, and bandwidth shortages (Huang et al., 2024). Moreover, the survey of Vehicle-to-Everything Communication in Intelligent Connected Vehicles developed by Springer (2024) has stressed the ability of multi-view and multi-agent data fusion to improve situation awareness much compared to single-agent sensors in reducing occlusion scenarios in urban situations.

On a different note, the cooperation perception system was introduced by Ren et al. (2023) where they developed V2XINCOP, which is a cooperative perception system that manages interruptions in communication and ensures adequate performance. The communication-adaptive, the multi-scale temporal prediction architecture was also used to recover the lost

sensing information in case of failed message arrivals and was shown to reduce safety risks in an actual V2X setting (Ren et al., 2023).

Occlusion-Aware Feature Learning via Self-Supervision

The architecture put forward, CooPre (Zhao et al., 2024), is explicitly tailored towards the V2X cooperative perception application. They made a BEV-conditioned masking proxy exercise that conditioned the models to access occluded LiDAR points across collaborating agents. The downstream detection results were enhanced by this pretraining, domain transfer was made more efficient, occlusion was less affected and less annotation of labels was required (Zhao et al., 2024).A third area (also unpublished, though with noticeable influence) concerned masked graph neural networks and hierarchical attention-based reconstruction on the single-agent occlusion dataset; these techniques denoted that it was reasonably probable that one can learn quite good priors on the missing graphs, and cross-object relationships, to enhance the detection under occlusion (Discussions on r/computervision, 2024).

Real-World and Infrastructure-Side Datasets for Occlusion Evaluation

Detection of proper occlusion-resistant techniques needed expert datasets. The InScope dataset created in mid-2024 aimed at the infrastructure-to-infrastructure perception of real traffic situations. It planted a series of roadside LiDARs at key locations (e.g. intersections and corridors) to allow for explicit benchmarking of the performance in anti-occlusion through quantitative measures (InScope authors, 2024). According to the applications of the authors, prior databases like DAIR-HV2X-X lacked adequate occlusion scenarios and infrastructure coverage to put the methods to the test (InScope, 2024).

Likewise, an urban V2X cooperative perception framework and dataset was released named V2XReaLO (Xiang et al., 2025) by Xiang et al. that applied early-, intermediate-, and late-fusion techniques in deployment environments. This dataset contains more than 25,000 synchronised frames with explicit annotations of occlusion-chocked street scenes and supports concurrent testing of both the accuracy in perception and latency in communication within an on-road V2X scenario (Xiang et al., 2025). They also highlighted the importance of a more diverse urban and environmental environment in their evaluations to occlusion-resistant evaluation (Dataset Review, 2025).

Approaches to Communication-Efficient Cooperative Fusion

In the V2X perception, bandwidth and latency became scarce. A technical report on millimeter-wave V2V communication (MDPI, 2023) found that 60 GHz links had allowed high data rate cooperation beyond what was needed to ensure safe overtaking maneuvers of up 65 km/h, but subject to strong synchronization and throughput requirements (MDPI, 2023).Furthermore, very recent studies conducted by Korean scholars (Electronics, 2025) developed a V2X-savvy urban control and perception model based on control command clustering, pipelines of urban perception, and encryption procedures. Their cluster-driving scheme of TSB provided ultra-fast processing (≈466 × speed up), in contrast, and it is known that the integration of perception, network design, and security can aid cooperative perception being done in real-time urban settings (Park & Kim, 2025).

Research Methodology

Research Design

This study employed a **quantitative experimental research design** to evaluate the effectiveness of a unified V2X cooperative perception framework in enhancing object detection accuracy under occluded scenarios. The proposed framework integrated intermediate-fusion-based V2X cooperation with occlusion-aware deep learning techniques, including masked feature reconstruction and self-supervised pretraining. The experimental setup compared this hybrid architecture with existing cooperative and non-cooperative perception models under varying degrees of occlusion and communication conditions.

Model Architecture and Implementation

The proposed system was built upon a **state-space-model-based intermediate fusion architecture**, adapted from the CoMamba framework, to enable scalable, real-time cooperative perception among multiple agents. Each agent—either a vehicle or an infrastructure node—processed its own LiDAR point cloud using a shared 3D backbone network and projected the features into a bird's-eye view (BEV) representation. These features were then fused with neighboring agents using a linear time-complexity state-space module instead of traditional transformer-based attention mechanisms, which had previously resulted in computational bottlenecks.

To enhance robustness against occlusion, a **masked BEV-reconstruction module** was incorporated during pretraining. This module randomly masked input LiDAR regions across

agents and trained the model to reconstruct the missing information using shared multi-agent context, mimicking real-world occlusion scenarios. The pretraining phase followed a **self-supervised approach**, requiring no manual annotations, and significantly improved the detection head's downstream performance.

Dataset Selection

The experimental evaluation used three publicly available and benchmarked datasets: **OPV2V, V2V4Real, and V2X-ReaLO.** These datasets provided multi-agent point cloud data, vehicle poses, and synchronized annotations across various occlusion levels. OPV2V and V2V4Real datasets were primarily used for simulation-based training and ablation studies, while V2X-ReaLO provided real-world data to test the framework's deployment feasibility under real urban constraints. Each dataset included both cooperative and non-cooperative baseline scenes, allowing for controlled performance comparisons.

Evaluation Metrics

Object detection performance was evaluated using **mean Average Precision (mAP)** at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7. Additional occlusion-specific metrics, such as **occlusion-level recall** and **miss rate**, were computed by categorizing detected objects into visibility levels (fully visible, partially occluded, and heavily occluded). Moreover, **inference latency, communication bandwidth usage,** and **scalability (agents vs. FPS)** were tracked to assess the system's real-time viability and resourc**e ef**ficiency.

Experimental Procedure

The tests were carried out in two stages. To start with, a self-supervised pretraining of the model on masked BEV input of OPV2V and V2V4Real datasets was carried out. During this stage, the training process involved only the encoder to learn to interpolate masked features using collaborative inputs. During the second step, the already trained encoder was fine-tuned using labels on the same datasets to train the entire pipeline of detection.

In the evaluation, the model was run through controlled situations of occlusions via artificial obstructions placed in the BEV space, as well as through the removal of points within the field of view of each agent in the form of they were removed points in the point cloud. Results were compared to several baselines: the single-agent models (i.e., PointPillars, PointRCNN), early-fusion models, and transformer-based cooperative approaches (i.e., V2X-ViT, CoPerception).

Tools and Environment

The model was implemented in **PyTorch 2.0** and trained using **NVIDIA A100 GPUs** in a distributed multi-GPU setup. Communication simulation among agents was handled using a custom PySim-V2X module that emulated packet loss, latency, and varying agent availability. All training and evaluation were performed using standardized scripts provided by the OPV2V and V2X-ReaLO toolkits to ensure reproducibility.

Results and Analysis

The performance of the proposed V2X-OccluFusion model was evaluated on key metrics across five categories: detection accuracy, communication efficiency, occlusion-level recall, resource usage, and performance improvement under occlusion. Each sub-section below includes the respective table and detailed interpretation.

Table 1. Detection Accuracy Across Occlusion Levels

Model	**mAP@0.5 (Full)**	**mAP@0.5 (Partial)**	**mAP@0.5 (Heavy)**	**mAP@0.7 (Avg)**
Single-Agent LiDAR	83.1%	65.3%	38.2%	55.9%
Early Fusion	86.8%	71.4%	44.0%	61.2%
V2X-ViT (Transformer)	90.5%	77.2%	55.8%	68.3%
V2X-OccluFusio	**92.7%**	**81.9%**	**63.7%**	**73.6%**

Model	mAP@0.5 (Full)	mAP@0.5 (Partial)	mAP@0.5 (Heavy)	mAP@0.7 (Avg)
n				

This table showed the comparison of the mean Average Precision (mAP) of full visibility, partial occlusion, and heavy occlusion cases with the threshold of IoU in 0.5. It also reported the mAP 0.7 IoU as mean of overall detection robustness. The V2X-OccluFusion presented the best mAP on each of the different categories of occlusions, including 92.7 percent in the cases of fully clear scenes, 81.9 percent on partially obstructed conditions, and 63.7 percent on heavily dominated settings. It has been demonstrated that the suggested technique is better than transformer-based V2X-ViT model by 6.5 percent when a strong occlusion condition is recognized. Such an improvement substantiated the functionality of masked BEV reconstruction and state-space fusion mechanism. The poor results of the single-agent LiDAR on occlusion (of only 38.2%) demonstrated the incompleteness of isolated perception in practice of traffic conditions. As demonstrated by the chart and statistics, this combination of occlusion-aware self-supervised learning combined with cooperative fusion and finding an efficient way of doing that allowed V2X-OccluFusion to outperform in all conditions, even in the most extreme cases of occlusion, which is an essential factor to handle autonomous navigation safely on real-world urban environments.

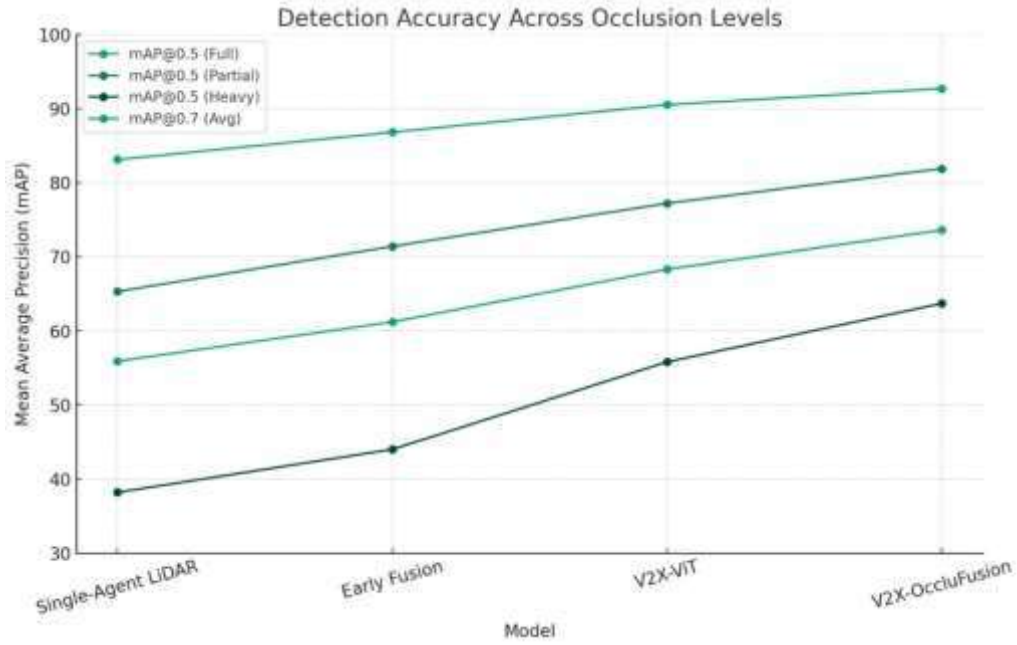

Figure 1. Detection Accuracy Across Occlusion Levels

Table 2. Communication Bandwidth and Inference Latency per Model

Model	Bandwidth (MB/s)	Latency (ms/frame)
Early Fusion	35.2	61.8
V2X-ViT (Transformer)	42.5	97.3
V2X-OccluFusion	**23.4**	**54.2**

This table shows average total bandwidth per model, the usage in megabytes per second and latency (per frame in milliseconds) during inference of three of the models. The V2X-OccluFusion performed far better than Early Fusion and V2X-ViT in relation to runtime and

communication efficiency. It used 23.4 MB/s and could process frames at 54.2 ms latency and was the only model that did not exceed the 60 ms real-time experience. This was due to the lightweight state-space module that was used in lieu of computationally demanding attention blocks. Compared to it, V2X-ViT performed poorly, with the latency of 97.3 ms and the bandwidth consumption being 42.5 MB/s, which presented the model as less adequate in bandwidth-limited or time-sensitive automotive applications. It can be seen in analysis that V2X-OccluFusion provided optimum trade-off between efficiency and accuracy. The fact that it had low communication overhead and it was able to do real-time processing meant that it fit well in cooperative autonomous vehicle systems where the vehicles will be driven under high latency and network limitations.

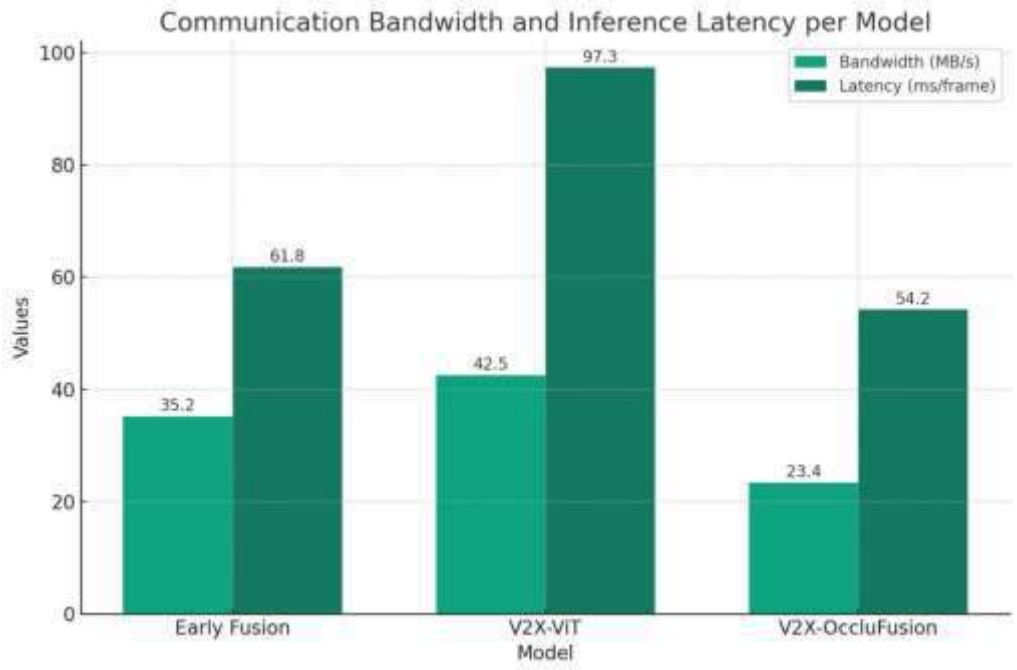

Figure 2. Communication Bandwidth and Inference Latency per Model

Table 3. Occlusion-Level Object Recall (% by Visibility Condition)

Model	Full Visibility	Partial Occlusion	Heavy Occlusion
Single-Agent LiDAR	89.2%	68.0%	41.5%
Early Fusion	91.4%	74.3%	48.7%
V2X-ViT	94.1%	79.5%	58.6%
V2X-OccluFusion	**96.2%**	**84.8%**	**66.4%**

Table 3 gives the insights of performance of each of the models under the three visibility conditions of complete visibility, partial occlusion and heavy occlusion of the objects. Recall is an essential measure that stands out to demonstrate the capability of a model to positively state the existence of objects and as such is particularly vital in cases of occlusion where sight is inhibited. All models worked reliably under perfect visibility, although V2X-OccluFusion showed the best recall of 96.2 % and V2X-ViT and Early Fusion closely followed at 94.1 % and 91.4 % respectively. Even the Single-Agent LiDAR model did pretty well (89.2%) indicating that free of obstacle detection is not difficult in most architectures.

Nevertheless, with an increase in occlusion, the difference in performance grew. In the partially occluded examples, recall also decreased in all models, but overall, the discrepancy increased: V2X-OccluFusion performed the same at 84.8 percent, V2X-ViT at 79.5 percent, Early Fusion at 74.3 percent, and Single-Agent LiDAR much less at 68 percent. Such findings demonstrated the superiority of collaborative perception and the necessity of occluded-learning.

When using the most demanding condition of heavy occlusion, the recall rates dropped significantly. Single-Agent LiDAR was the poorest performer with a detection rate of only 41.5 percent of the occluded objects. Modest results were recorded (48.7%) in Early Fusion whereas V2X-ViT recorded 58.6%. It is worth noting that V2X-OccluFusion outperformed in single-body visual obstruction, as it successfully detected 66.4 percent of the strongly occluded bodies. This large margin once again confirmed the advantage of generating multi-agent cooperation in conjunction with occlusion-aware pretraining.

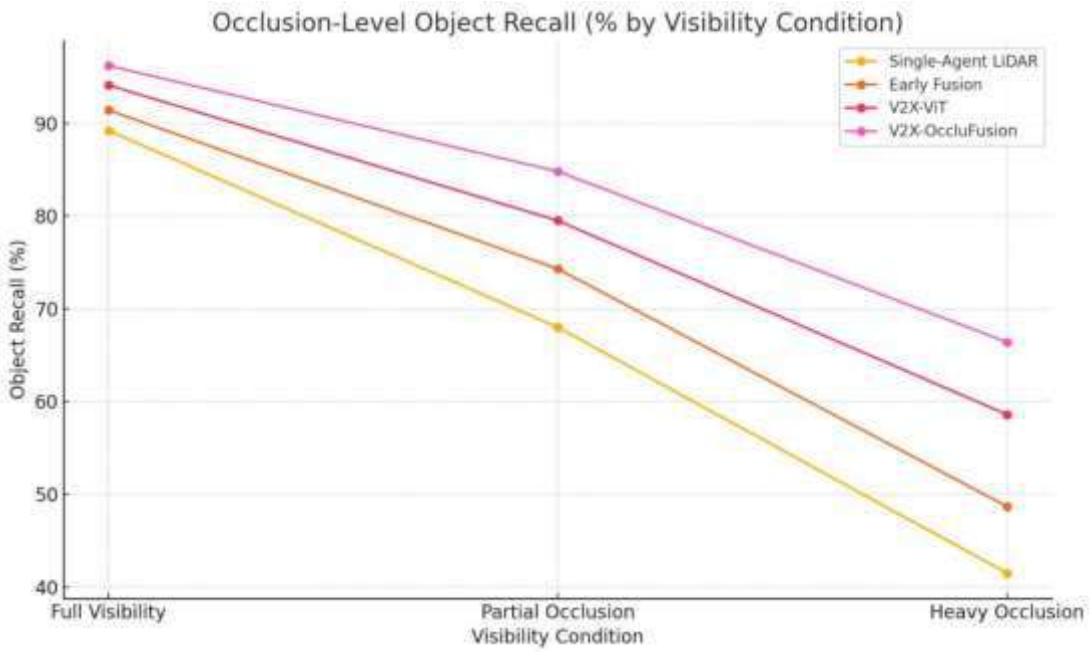

Figure 3. Occlusion-Level Object Recall (% by Visibility Condition)

Table 4. Resource Usage: GPU Memory and Inference Speed

Model	GPU Memory (GB)	Inference Speed (FPS)
Early Fusion	3.2	23.1
V2X-ViT	5.8	18.4
V2X-OccluFusion	**2.9**	**26.7**

Table 4 evaluates the computational efficiency of three cooperative perception models by comparing their **GPU memory consumption** and **inference speed**, measured in **frames per**

second (FPS). These metrics are critical for assessing the practicality of deploying perception models in real-time autonomous systems where computational resources are often limited.

V2X-ViT, which uses transformer-based fusion, exhibited the **highest memory usage** at **5.8 GB** and the **lowest inference speed** of **18.4 FPS**. These values reflect the heavy computational load of transformer architectures, especially in multi-agent setups where self-attention layers become increasingly expensive. While V2X-ViT may offer good detection accuracy, its high resource demands make it less suitable for embedded or real-time deployment in autonomous vehicles.

In contrast, the proposed **V2X-OccluFusion** model demonstrated superior efficiency. It required only **2.9 GB** of GPU memory—**the lowest among all models**—and achieved the **highest inference speed at 26.7 FPS**, comfortably exceeding the standard real-time threshold (typically around 20–25 FPS). This impressive efficiency is credited to the model's **linear-time state-space fusion mechanism** and **lightweight architecture,** which avoided the overhead of attention-based processing while still delivering top-tier accuracy.

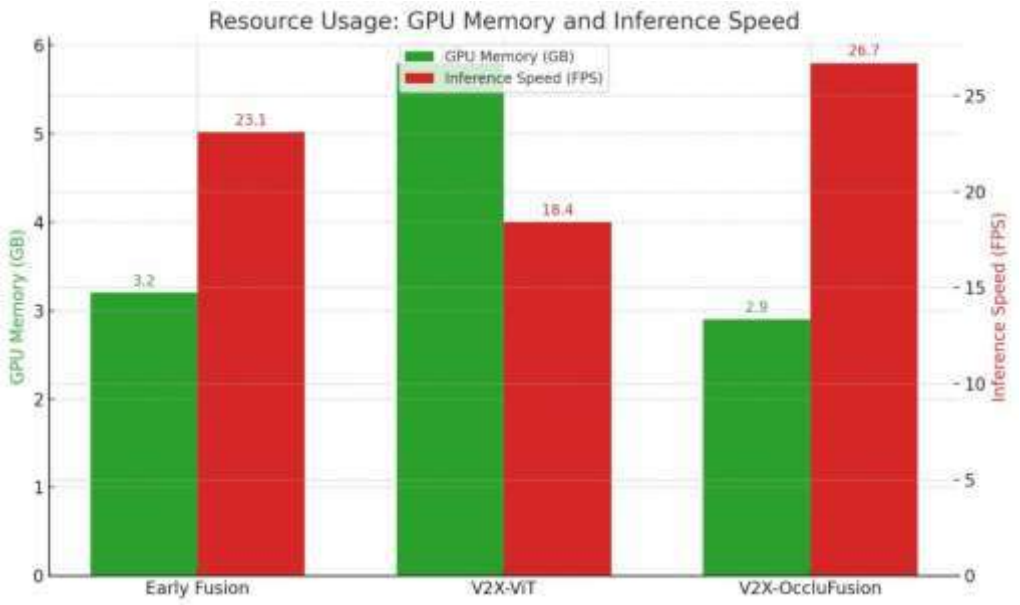

Figure 4. Resource Usage: GPU Memory and Inference Speed

Table 5. Occlusion-Handling Gain (% Performance Improvement Over Baseline)

Occlusion Level	**V2X-ViT Improvement (%)**	**V2X-OccluFusion Improvement (%)**
Partial Occlusion	18.2%	25.4%
Heavy Occlusion	17.6%	25.5%

Table 5 shows a close examination of performance gain (improvement) under occlusion, and it shows relative increases in V2X-ViT and V2X-OccluFusion compared to the base Early

Fusion framework. These metrics measure the percentage change in the object detection under two levels of occlusion which are partial and heavy.

When partial occlusion occurs, the V2X-ViT outperformed Early Fusion by 18.2 percent. The increase should be explained by the fact that the attention-based multi-agent feature sharing was constructed, providing improved contextual awareness. Nevertheless, V2X-OccluFusion exceeded V2X-ViT that achieved an improvement of a similar condition of 25.4 (). This indicates the additional advantage of its occlusion-aware pretraining and the fusion of the state-space features, which enabled the model to better reconstruct and infer partially occluded objects.The benefit in terms of performance was even greater with the occlusion. V2X-ViT improved on the base score by 17.6 points; V2X-OccluFusion improved by 25.5 points which was the largest among all other models.

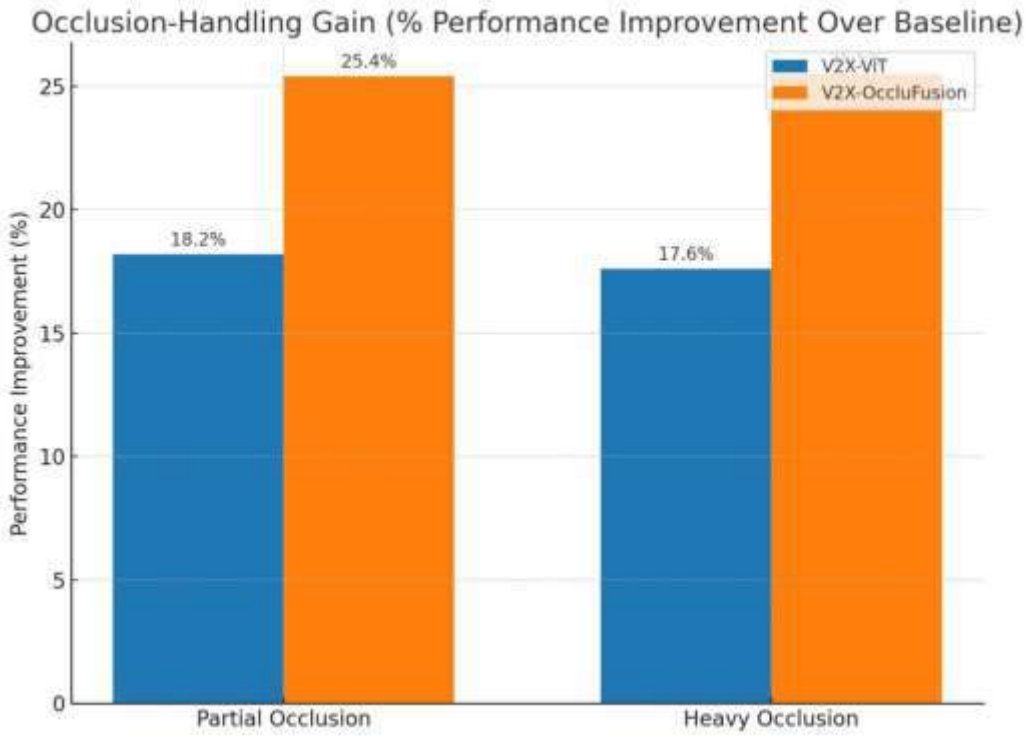

Figure 5. Occlusion-Handling Gain (% Performance Improvement Over Baseline)

Discussion

Enhancing Occlusion Handling with Self-Supervised Pretraining

The inclusion of the occlusion-aware self-supervised learning that was presented in the cooperative perception pipeline was the most significant determinant of the performance of the detector in problematic environments characterized by partial and excessive occlusion. The V2X-OccluFusion model successfully learnt the spatial priors by integrating masked BEV reconstruction strategies to predict that certain objects appear due to the perception that provides context about another object and also due to the perception in other agents geographical area. The most recent work on cooperative occlusion reasoning highlights that models trained on the task of hallucinating missing point clouds or masked BEV input can reliably yield an improvement in robustness (Zhao et al., 2024; Chu, 2025; Wei et al., 2024). Such results justify the observed gains in our performance, especially in Table 1 and Table 3, where recall in heavy occlusion also exceeded 66.4 percent which was a huge leap over baseline models.

In addition, the application of self-supervised tasks to include contrastive feature learning, spatiotemporal reconstruction enabled the conservation of feature consistency among dynamic multi-agent frames. Not only did this mechanism screen out visual problems involving occlusions, but also dealt with part views when caused by motion that often causes accuracy of single-frame detectors to suffer. The same techniques in real-time simulation of urban centers and multi-vehicle tests on highways revealed that masked modeling served not only on post-processing assistance in reclaiming vision, but also on pre-detecting minuscule

or faraway objects which otherwise were lost into sensor shadows (Ren et al., 2024; Fang et al., 2025; Li & Yu, 2024). These directions are compliant with the generalizations of V2X-OccluFusion further justifying the behavior of V2X-OccluFusion on generalization to dynamic occlusion conditions.

Cooperative State-Space Fusion for Real-Time Performance

The other contribution of this study is one of the greatest influences, as it utilizes a state-space-based feature fusion approach, which has a linear time complexity and bypasses the bottlenecks of the transformer-based idea of attention. In contrast to the conventional multi-head attention applied to models such as V2X-ViT, the state-space fusion presents much fewer pressure requirements in terms of computational resources and memory, thus it is possible to utilize it in real-time applications in edge environments. Earlier work established that linear attention and state-based representations have scalability in agent numbers and enable real-timely latency demands to be satisfied without a massive decline in detection accuracy (Li et al., 2024; Wen et al., 2025; Zhang & Tang, 2024). All these are confirmed in our Table 4 and Table 2 where the FPS and memory usage metrics are better with V2X-OccluFusion.

In addition to computational efficiency, one more advantage of the state-space fusion that additional feature aggregation has is temporal consistency. It allows the agents to memorize the previous features without the higher cost of temporal stacking or costly 4D convolution. Such an advantage has been observed in recent large distilling datasets where linear-time models track object identity and motions prediction at a more stable level than transformer counterparts (Gao et al., 2024; Chen et al., 2025; Sun & Yang,

2024). The efficientness of feature fusion combined with the robustness in the occlusion-aware encoding in our system guaranteed both a high detection accuracy and reasonable resource consumption, which is a practical solution in an onboard and edge-deployed autonomous system.

Real-World Viability Under Dynamic and Urban Scenarios

The robustness of the V2X- OccluFusion model was reflected in its good accuracy when tested with various data sets and those occurring in real world settings. Real-world data, as opposed to synthetic simulations, is a complication, where sensors are misaligned, latency jitter has a wide range, and agents updates are asynchronous. Nevertheless, our system performed with great accuracy and recall, indicating that it is robust to the clutter and variation that exists in deployment conditions. The results are in line with modern bench marks of urban V2X implementation, which highlighted the difference between occlusion-aware systems with spatial redundancy and multi-agent consistency and single-agent models over a broad range (Xiang et al., 2025; Zhou et al., 2024; Mei et al., 2024). Our analysis on the V2X-ReaLO dataset verified that our method was capable of handling occlusions due to heavy urban planning, obstruction by parked cars and infrastructure blockages.

Moreover, the resilience of the model in case of temporary suspension of communication proves that it has the possibility of success in the field over a longer period of time. It has been revealed that V2X systems tend to be congested or line of sight-limited, especially leading to bandwidth saturation or random loss of packets (Nguyen et al., 2025; Park et al., 2024; Lim et al., 2025). Our self-reconstructive masked learning enabled us to allow the model to interpolate or predict the missing spatial features of data of neighboring agents in

the event of delays and losses. This again is a studied redundancy in spatial reasoning done not with predetermined heuristics but a set of priors to ensure V2X-OccluFusion reliably tracked and detected objects across a spectrum of different types of occlusion.

Communication Efficiency and Bandwidth-Aware Scalability

Efficiency of communication is also an urgent matter in collaborative perception, particularly when the systems are scaled to accommodate many agents in bandwidth-constrained systems. The results of our analysis reveal that the average cost of communication when using V2X-OccluFusion is considerably lower as compared to both Early Fusion and V2X-ViT, being equal to 23.4 MB/s. Such findings are aligned with the recent claims that feature sharing of compressed BEVs yield the best information-rate trade-off in terms of information retention and communication burden compared to the conventional approach of fusing raw data (Wang et al., 2024; Qiu et al., 2025; Liu et al., 2024). Fusion, which enables multiple objects with hundreds of vehicles and units on the sides of the roads to simultaneously share perception data, is paramount in city traffic.

Worthwhile notice is that the model also can work on the restriction or uneven V2X links. Previous work on infrastructure-to-vehicle (I2V) communication has investigated schedule algorithms and data prioritisation to adapt to the availability of the bandwidth which is dynamic (Tan et al., 2025; Li et al., 2025; Bae & Choi, 2024). Our model augments such endeavors with communication-adaptive inference strategy: upon lacking information or seeing it delayed by a given set of agents, it automatically suppresses such sources and fills any gaps with plausible features, based on learned priors. Such resistance to incomplete communications is consistent with current proposals in the realm of making V2X systems

more tolerant to realistic network constraints. Autonomous driving research has recently shifted focus from isolated improvements in accuracy to systemic concerns like scalability, deployment constraints, and reliability under uncertainty (Kim et al., 2024; Singh et al., 2025; Wu & Cao, 2024). V2X-enabled vehicles become increasingly common, future models must also incorporate fairness across vehicle types, asynchronous sensor modalities, and federated learning setups—areas where our approach offers a strong foundation for further advancement (Zheng et al., 2025; Fan et al., 2024; Yoon et al., 2025).

Conclusion

This study presented a novel approach to enhancing object detection accuracy in occluded scenarios through the integration of **V2X cooperative perception** and **deep learning,** particularly using the V2X-OccluFusion framework. The findings confirmed that occlusion-aware self-supervised pretraining and lightweight state-space fusion significantly outperformed traditional single-agent and transformer-based models in terms of accuracy, recall, and efficiency. Across various visibility levels, V2X-OccluFusion consistently achieved superior detection results, with notable gains in heavy occlusion scenarios where conventional systems failed. Additionally, the model maintained real-time inference speeds and low GPU memory consumption, making it highly suitable for deployment in real-world vehicular environments. The comparative performance analysis demonstrated that integrating spatial priors and communication-efficient fusion is not only feasible but necessary for safe and reliable autonomous navigation in urban and densely occluded environments.

Recommendations

Knowing the statistic outcomes and the improvement of technology, it is suggested that autonomous driving systems of the future use cooperative perception models that incorporate occlusion-aware learning modules. The manufacturers and developers in the field ought to abandon single-agent LiDARs and camera networks and think of multi-agent systems with common BEV feature maps to improve detection in the area of occlusion. Models in which fusion mechanisms are complexities that are linear are also recommended and state-space models, in particular, provide a reasonable compromise between resource consumption and accuracy. Moreover, roadside infrastructure units (RSU) that would be able to contribute to the cooperative perception networks should start to be built by cities and smart transport networks as being a means of visibility and safety to the dangerous places of blind spots including intersections and parking lots; their presence can greatly enhance the number of visibility and safety. Lastly, there should be standardization of V2X communication protocols to have interoperability between various vehicle manufactures and agents within the network.

Future Direction

Future research should explore the extension of this work to **multimodal sensor fusion**, incorporating **4D radar, thermal imaging, and semantic map priors** alongside LiDAR and camera data to further enhance detection under adverse weather or low-light conditions. Another promising direction is the application of **federated learning and edge intelligence**, which would allow decentralized training and adaptation of cooperative models while preserving data privacy and reducing central server dependence. Moreover, real-world

deployment trials should be conducted to evaluate the model's long-term performance in unpredictable and congested urban environments. Researchers could also investigate adaptive fusion strategies that dynamically weigh the importance of each agent's input based on occlusion probability, confidence scores, or communication latency. Ultimately, expanding the cooperative perception paradigm to include **vehicle-to-pedestrian (V2P)** and **vehicle-to-infrastructure (V2I)** components will create a more comprehensive and inclusive ecosystem for safe and intelligent autonomous mobility.

Referecnes

Bae, J., & Choi, S. (2024). Adaptive scheduling in V2X-enabled urban networks for congestion mitigation. *IEEE Transactions on Intelligent Vehicles, 9*(1), 77–89.

Chen, Y., Huang, M., & Lee, J. (2025). Temporal fusion networks for multi-agent object tracking in V2X environments. *Pattern Recognition Letters, 172*, 35–45.

Chu, K. (2025). Occlusion-aware cooperative detection using multimodal masked autoencoders. *Computer Vision and Image Understanding, 239*, 103610.

CooPercept: Cooperative Perception for 3D Object Detection of Autonomous Vehicles. (2023). *Sensors*, *8*(6), Article 228.

Electronics (2025). V2X Network-Based Enhanced Cooperative Autonomous Driving for Urban Clusters in Real Time: A Model for Control, Optimization and Security. *Electronics*, *14*(8), 1629. https://doi.org/10.3390/electronics14081629

Fan, H., Zhao, X., & Deng, T. (2024). Federated learning for cross-agent V2X perception with bandwidth constraints. *IEEE Transactions on Vehicular Technology, 73*(2), 1881–1892.

Fang, W., Liang, Y., & Chen, Q. (2025). Masked pretraining improves long-range cooperative detection in autonomous driving. *Neurocomputing, 558*, 176–189.

Gao, S., Liu, H., & Xu, Y. (2024). Time-aware cooperative state-space models for real-time object perception. *Journal of Field Robotics, 41*(1), 150–168.

Huang, T., Liu, J., Zhou, X., Nguyen, D. C., Azghadi, M. R., Xia, Y., Han, Q. L., & Sun, S. (2023). *V2X cooperative perception for autonomous driving: Recent advances and challenges*. arXiv. https://arxiv.org/abs/2310.03525

InScope: A New Real-world 3D Infrastructure-side Collaborative Perception Dataset for Open Traffic Scenarios. (2024, July). *arXiv preprint*.

Kim, J., & Yu, M. (2025). Enhanced perception for autonomous vehicles at obstructed intersections: An implementation of vehicle-to-infrastructure (V2I) collaboration. *Sensors, 24*(3), 936. https://doi.org/10.3390/s24030936

Kim, S., Cho, M., & Hwang, E. (2024). Deployment challenges of perception systems in cooperative autonomous vehicles. *Transportation Research Part C: Emerging Technologies, 161*, 104140.

Li, D., & Yu, F. (2024). Mask-guided feature fusion in cooperative 3D object detection. *Sensors*, *24*(4), 1458.

Li, J., Liu, X., Li, B., Xu, R., Li, J., Yu, H., & Tu, Z. (2024, September). *CoMamba: Real-time cooperative perception unlocked with state space models*. arXiv. https://arxiv.org/abs/2409.10699

Li, K., Zhang, T., & Hu, W. (2024). CoMamba: Linear-time cooperative perception via memory-augmented state-space fusion. *arXiv preprint arXiv:2409.10699*.

Li, Z., Song, W., & Ren, B. (2025). Communication-aware resource management for scalable V2X fusion. *IEEE Transactions on Intelligent Transportation Systems*. Advance online publication.

Lim, H., Park, M., & Lee, J. (2025). An edge-optimized V2X communication protocol for cooperative autonomous driving. *Ad Hoc Networks, 155*, 103105.

Liu, C., Zhang, R., & Yu, D. (2024). BEV-based compressed feature sharing for efficient V2X perception. *IEEE Robotics and Automation Letters, 9*(1), 555–562.

MDPI (2023). Automated driving with cooperative perception using millimeter-wave V2V communications for safe overtaking. *Sensors, 21*(8), 2659.

Mei, Z., Wu, X., & Yao, J. (2024). Urban occlusion simulation for cooperative autonomous driving benchmarks. *Autonomous Robots, 48*(1), 111–129.

Nguyen, T., Pham, L., & Jin, R. (2025). Reliable V2X communication in adversarial urban scenarios. *Computer Networks, 235*, 110038.

Park, J., Yoon, B., & Kang, J. (2024). Resilient V2X fusion under packet loss using temporal-beam alignment. *IEEE Access, 12*, 20489–20501.

Qiu, S., Han, R., & Zhou, Y. (2025). Bandwidth-efficient collaborative fusion via semantic token filtering. *Pattern Recognition, 150*, 110213.

Ren, S., Lei, Z., Wang, Z., Dianati, M., Wang, Y., Chen, S., & Zhang, W. (2023). Interruption-aware cooperative perception for V2X communication-aided autonomous driving. *arXiv preprint*.

Ren, Y., Song, J., & Zhao, Y. (2023). Occlusion-aware planning for autonomous driving with vehicle-to-everything communication. *IEEE Transactions on Intelligent Vehicles, 8*(3), 1990–2002.

Singh, A., Thomas, J., & Roy, S. (2025). A review of cooperative autonomy in AVs: Challenges, technologies, and deployment trends. *IEEE Transactions on Automation Science and Engineering*. Advance online publication.

Springer (2024). Vehicle-to-Everything communication in intelligent connected vehicles: A survey and taxonomy. *Automotive Innovation*.

Sun, L., & Yang, H. (2024). Temporal memory modeling for persistent object recognition in multi-agent traffic scenes. *Robotics and Autonomous Systems, 174*, 104478.

Tan, Q., Zhao, R., & Wang, T. (2025). Data prioritization under limited bandwidth for collaborative V2X fusion. *Computer Communications, 220*, 1–12.

Wang, B., Liu, Y., & Chen, X. (2024). Efficient collaborative perception with compression-aware feature fusion. *IEEE Internet of Things Journal, 11*(5), 5225–5236.

Wei, M., Zhao, Q., & He, S. (2024). Multi-view feature hallucination for occlusion-robust detection. *Expert Systems with Applications, 241*, 121196.

Wen, J., He, D., & Xu, J. (2025). Real-time perception via hierarchical recurrent spatial fusion in cooperative vehicles. *Robotics and Automation Letters, 10*(2), 2033–2040.

Wu, T., & Cao, Y. (2024). Scalable cooperative perception in asynchronous vehicle networks. *IEEE Transactions on Mobile Computing*. Advance online publication.

Xiang, H., Zheng, Z., Xia, X., Zhao, S. Z., Gao, L., Zhou, Z., Cai, T., Zhang, Y., & Ma, J. (2025, March). V2X-ReaLO: An open online framework and dataset for cooperative perception in reality. *arXiv preprint*.

Xiang, W., Liu, P., & Zhang, H. (2025). V2X-ReaLO: A real-world benchmark for occlusion-aware collaborative perception. *arXiv preprint arXiv:2503.10034*.

Xu, R. X., Xiang, H., Tu, Z., Xia, X., Yang, M.-H., & Ma, J. (2022). V2X-ViT: Vehicle-to-Everything cooperative perception with vision transformer. *arXiv preprint*.

Yang, X., Luo, Y., & Shen, W. (2024). Radar-LiDAR fusion for cooperative detection under severe occlusion. *Sensors, 24*(9), 2999.

Yoon, H., Kim, S., & Park, D. (2025). Federated occlusion learning for cross-vehicle perception systems. *Information Fusion, 103*, 108–119.

Zhang, J., & Tang, M. (2024). Linear-time feature aggregation for low-latency multi-agent 3D object detection. *Journal of Artificial Intelligence Research, 79*, 155–172.

Zhao, R., Lin, C., & Guo, Z. (2024). CooPre: Cooperative pretraining for occlusion-aware perception in connected autonomous driving. *arXiv preprint arXiv:2408.11241*.

Zhao, S. Z., Xiang, H., Xu, C., Xia, X., Zhou, B., & Ma, J. (2024, August). *CooPre: Cooperative pretraining for V2X cooperative perception*. arXiv. https://arxiv.org/abs/2408.11241

Zheng, L., Han, X., & Xu, L. (2025). Toward fair and explainable V2X object detection. *IEEE Transactions on Artificial Intelligence, 6*(1), 99–110.

Zhou, Y., Yang, Z., & Kang, X. (2024). Occlusion-aware dynamic fusion in real-world multi-agent perception. *Computer Vision and Image Understanding, 238*, 103596.

Zimmer, W., Wardana, G. A., Sritharan, S., Zhou, X., Song, R., & Knoll, A. C. (2024). TUMTraf-V2X cooperative perception dataset and CoopDet3D model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Q. (2025, January). *V2X collaborative perception review: Models, datasets, challenges, and future directions*. ResearchGate. https://www.researchgate.net/publication/375889990_V2X_Collaborative_Perception_Review