# Mapcooper: A Communication-Efficient Collaborative Perception Framework via Map Alignment

Huan Qiu[1], Kai Liu[2], Bijun Li[1], Youchen Tang[1], Jinsheng Xiao[2], Jian Zhou[1*]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China-(huanqiu, lee, youchentang, jianzhou)@whu.edu.cn
[2] Electronic Information School, Wuhan University, China-(kailiu, xiaojs)@whu.edu.cn
* Correspondence: jianzhou@whu.edu.cn

**Keywords:** V2X, Cooperative perception, HD map, Autonomous driving

**Abstract**

V2I collaborative perception improves awareness of the dynamic driving environment by exchanging multi-viewpoint information through communication, establishing itself as a key element of intelligent transportation systems. Despite its advantages, this method requires a balance between communication bandwidth and perception performance. To address this challenge, we propose a map-mask designed to align with perceptual spatial features, enabling precise background filtering to isolate critical areas for communication. During the sender's compression phase, the map-mask filters out background elements and extracts key features from critical areas, significantly reducing communication bandwidth consumption. During the receiver's decompression phase, the map-mask restores scene context and enhances spatial information surrounding critical areas, ensuring the preservation of perception performance. Based on this map alignment, we develop Mapcooper, a unified collaborative perception framework that optimizes the balance between communication bandwidth and perception performance. We evaluated Mapcooper's effectiveness via extensive experimentation using the large-scale V2X-Seq-SPD dataset. The results demonstrate that Mapcooper outperforms existing collaborative perception approaches with respect to perceptual accuracy while minimizing communication transmission costs.

## 1. Introduction

Accurately perceiving the complex driving environment is crucial for ensuring the safe and reliable operation of autonomous vehicles (AVs). Driven by the rapid advancements in deep learning technologies, single-vehicle perception systems have demonstrated significant improvements in tasks such as object detection(Lang et al., 2019, Yin et al., 2021), semantic segmentation(Yan et al., 2022, Zeng et al., 2024), and depth estimation(Chuah et al., 2022, Cheng et al., 2024). Despite these advancements, single-vehicle systems still suffer from inherent limitations. Challenges such as occlusions from obstacles, sparse sensor data at longer distances, and limited field of view constrain their ability to perceive the environment accurately and comprehensively. These constraints arise from the fundamental nature of single-vehicle systems, which are limited to capturing the environment from a singular, frequently obstructed viewpoint.

To address these limitations, Vehicle-to-Infrastructure (V2I) collaborative perception has emerged as a viable solution(Yu et al., 2024, Xu et al., 2022b), offering the advantage of fusing information from multiple viewpoints to enhance environmental perception. V2I systems leverage real-time data exchange between vehicles and infrastructure to extend the visibility range and enhance situational awareness. However, the primary challenge in implementing V2I perception lies in striking an optimal balance between high perception performance and efficient communication bandwidth usage. Transmitting raw sensor data across a network in real time demands substantial bandwidth, which not only risks communication bottlenecks but also increases the likelihood of delayed information exchange, potentially compromising safety. Recent studies(Liu et al., 2020, Wang et al., 2020, Wang et al., 2023) have explored various strategies to mitigate this issue by compressing the transmitted data while retaining essential perceptual information. Traditional approaches (Liu et al., 2020, Wang et al., 2020, Li et al., 2021)often focus on reducing the size of global feature maps; however, this can lead to inefficiencies, as much of the transmitted data might not be directly relevant to the perception task. As a result, these methods can inadvertently consume excessive bandwidth while offering only marginal performance gains. In response, contemporary approaches(Yu et al., 2024, Hu et al., 2022, Wang et al., 2023) have shifted towards more selective filtering mechanisms, where only the most relevant perceptual features from critical regions are identified and transmitted. Some of these methods employ mathematical models to compress high-dimensional data from targeted areas, yielding reductions in bandwidth usage. Despite their successes, these approaches still face challenges in accurately isolating and compressing features from areas of high perceptual importance, limiting their overall effectiveness in achieving both minimal communication bandwidth and optimal perception performance.

To address these challenges, we propose a novel communication compression and decompression strategy guided by map-masks, which leverages high-precision map data to improve efficiency. In the sender's compression phase, the map-mask serves as an intelligent filter, dynamically selecting critical areas based on real-time environmental information and the spatial distribution of key dynamic target features. This selective approach optimizes the transmission of relevant data while minimizing unnecessary communication overhead, thereby significantly conserving communication bandwidth. In the receiver's decompression phase, the map-mask not only restores critical areas but also plays a pivotal role in reconstructing the surrounding scene context. By leveraging spatial and semantic

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-673-2025 | © Author(s) 2025. CC BY 4.0 License.

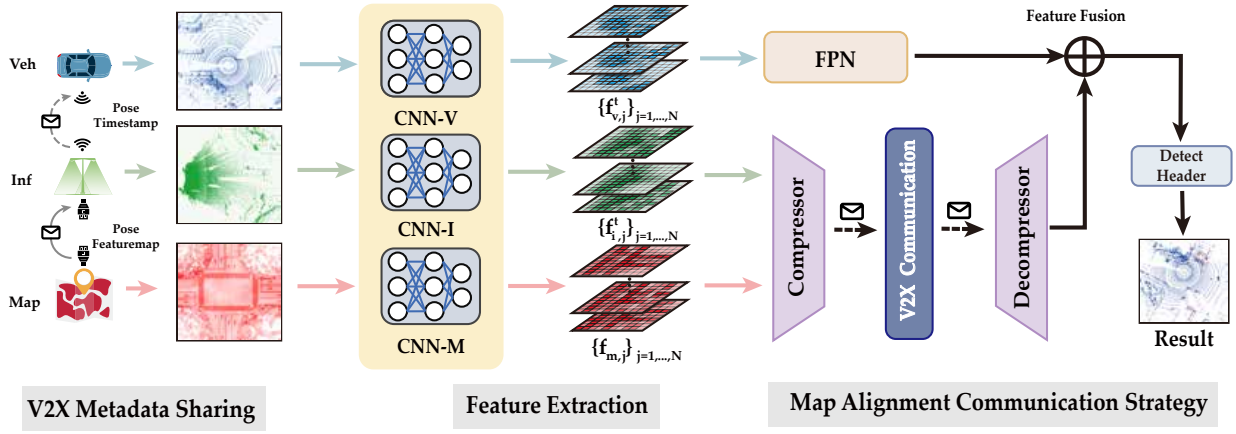



Figure 1. Overall framework of Mapcooper. It is composed of three consecutive stages: V2X metadata sharing, feature extraction,and map alignment communication strategy (MACS).

relationships encoded within the high-precision map, the system enriches the perception of the environment, ensuring accurate representation of key features such as object boundaries, spatial relationships, and dynamic movements This process preserves perceptual quality and mitigates potential information loss that may occur due to aggressive data compression, maintaining perception performance.

In this paper, we introduce Mapcooper, a novel collaborative perception framework designed to strike an optimal balance between communication bandwidth and perception performance. The system architecture is illustrated in Fig.1, we propose a map alignment communication strategy(MACS), where critical perceptual features are selected during the compression phase to conserve communication bandwidth, and scene context is restored during decompression to maintain perception performance. To evaluate the effectiveness of Mapcooper, we performed extensive experimentation on the large-scale V2X-Seq-SPD(Yu et al., 2023) dataset. The results demonstrate that Mapcooper outperforms state-of-the-art collaborative perception methods, delivering superior perception performance while minimizing communication transmission costs.

## 2. Related work

### 2.1 Collaborative Perception

Collaborative perception aims to enhance autonomous systems by enabling information sharing and fusion across multiple agents, such as vehicles and infrastructure, to improve overall perception accuracy and robustness. The efficiency of such systems largely depends on the strategy employed for message sharing, which can be broadly categorized into early, intermediate, and late fusion. Early fusion(Chen et al., 2019b), where raw sensor data is shared between agents, offers rich information but imposes high bandwidth demands, making it less practical for real-time applications. Late fusion(Rawashdeh and Wang, 2018), which aggregates final detection results like bounding boxes or object classifications, typically reduces communication costs but lacks the contextual depth needed for complex scenarios. As a middle ground, intermediate fusion has emerged as a promising approach, where intermediate feature representations from neural networks are exchanged. This method strikes a balance between preserving rich perceptual information and minimizing transmission bandwidth. Most recent methods in collaborative perception have focused on intermediate fusion strategies to balance perception accuracy and communication bandwidth. DiscoNet(Li et al., 2021), for example, leverages knowledge distillation to align feature representations by constraining the student model to learn from a teacher model based on raw data fusion. F-Cooper(Chen et al., 2019a) introduced one of the earliest feature-level collaboration methods, employing a max-based function to equally weight interaction information between agents. V2VNet (Wang et al., 2020)further refines feature exchange by implementing a spatially-aware message passing mechanism, where weights are adaptively assigned to agents based on their spatial positions. When2com(Liu et al., 2020) utilizes an attention mechanism to dynamically adjust communication groups, optimizing bandwidth usage by selectively transmitting the most relevant features. Similarly, Where2comm(Hu et al., 2022) capitalizes on the sparsity of foreground information in detection tasks, reducing the communication load by prioritizing essential features. More recently, V2X-ViT(Xu et al., 2022b) has proposed a Transformer-based framework that accounts for the heterogeneity across V2X systems, unifying the fusion process while considering the diversity of data sources. Lastly, CoBEVT(Xu et al., 2022a) integrates multi-camera inputs to generate BEV map predictions through feature-level collaboration, further highlighting the trend toward optimizing multi-modal data fusion in collaborative perception systems.

### 2.2 LiDAR-Map Fusion Perception

LiDAR-map fusion for 3D object detection has emerged as a promising direction due to its ability to incorporate rich prior knowledge from high-definition (HD) maps, which enhances detection accuracy in complex environments. While most LiDAR-based methods primarily rely on integrating data from cameras and radars, the fusion of HD maps introduces valuable contextual information. Early work such as HDNet(Yang et al., 2018) represented HD maps through rasterization and concatenated them with LiDAR features in the bird’s-eye view (BEV), demonstrating the potential of map-based fusion in 3D perception. MapFusion(Fang et al., 2021) further refined this approach by employing a 2D feature extractor for HD maps and concatenating these features with those learned from a modern LiDAR-based detector. This method highlighted the benefits of integrating map and sensor data, though primarily at the feature level. Additionally, LaneFusion(Fujimoto et al., 2022) explored the use of lane maps along with LiDAR point clouds, improv-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-673-2025 | © Author(s) 2025. CC BY 4.0 License.

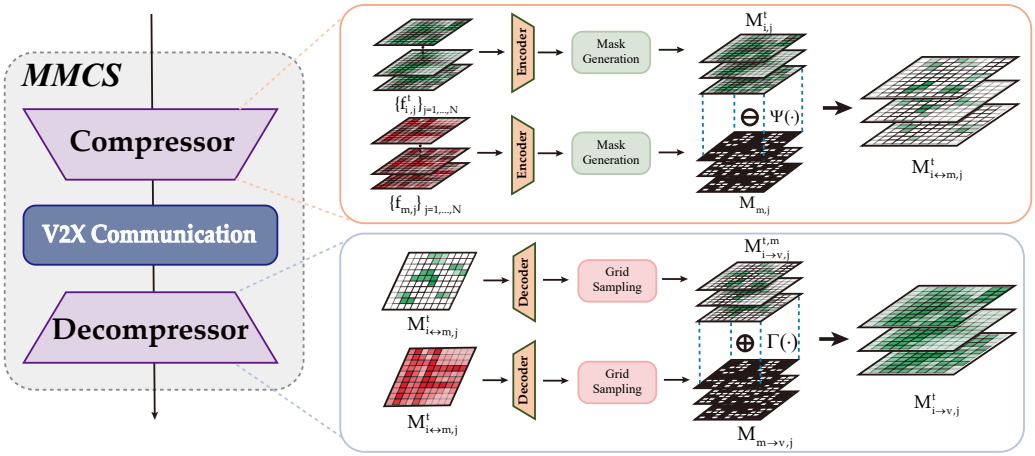



Figure 2. Architecture of the map alignment communication strategy (MACS)

ing orientation prediction by incorporating geometric information. However, recent developments such as MENet (Huang et al., 2023)have tested several attention mechanisms as fusion modules, revealing that dilated convolution is particularly effective in encoding HD maps, offering more nuanced feature extraction capabilities. Additionally, MIM(Xiao et al., 2024) utilizes an attention mechanism to effectively align multi-view BEV features with HD map features, enabling efficient cross-modal feature fusion between maps and images. Despite these advancements, LiDAR-map fusion methods have yet to see extensive application in the V2X domain, where the fusion of HD maps with real-time vehicle and infrastructure sensor data remains largely unexplored. The potential to leverage HD maps in vehicle-to-everything (V2X) communication could provide significant benefits by enhancing situational awareness and improving collaborative perception in connected vehicle systems.

## 3. Method

Overall framework of Mapcooper is illustrated in Fig.1, which includes three key components: V2X metadata sharing, feature extraction and map alignment communication strategy (MACS).In this section, we will provide a detailed introduction to the proposed collaborative perception network and its related technical modules.

### 3.1 V2X Metadata Sharing

In the entire process of vehicle-to-infrastructure (V2I) collaborative perception, the metadata sharing strategy serves as the foundation for achieving cooperative perception functionality. Effective information transmission between participating agents (vehicles, infrastructure, and map servers) is critical to the performance and accuracy of the perception system. The goal of metadata sharing is to ensure that all agents can obtain each other's key positional information, sensor data, and other useful perception data under synchronized timestamps. As intelligent transportation systems continue to evolve, metadata sharing via V2I has become a core component in vehicle-road collaborative perception systems. Therefore, designing a rational data-sharing strategy that ensures the synchronization and accuracy of information transmission provides a solid foundation for subsequent feature extraction and processing stages. At the initial stage of V2I collaborative perception, each agent shares metadata such as poses, timestamps, and sensor states within the communication network. The vehicle side communicates with the infrastructure side through a wireless network, while the infrastructure side, vehicle side, and map server communicate via Ethernet connections. The map mask feature map is stored offline on both the vehicle and roadside devices, with periodic updates delivered through the cloud. In this work, we assume that wireless communication transmission for metadata achieves good synchronization, meaning that each agent can receive pose information at synchronized timestamps to perform collaborative LiDAR point cloud calibration between different agents. Given the variety of sensors involved, the metadata being shared includes not only positional information but also details about the sensor configurations, such as intrinsic and extrinsic calibration parameters and time synchronization data. For agent $j$ at time $t$, the shared metadata can be represented as:

$$M_j^t = \left\{P_j^t, T_j^t, S_j^t\right\} \tag{1}$$

where $P_j^t$ represents the position and orientation of the agent, $T_j^t$ the timestamp, and $S_j^t$ the sensor configuration details. One of the key challenges in V2I metadata sharing is ensuring synchronization across multiple agents. All metadata must be shared in real-time, with minimal delay, to guarantee that all agents are working with temporally consistent information. In this paper, we assume that the synchronization mechanism is robust and that metadata from all agents can be shared within a sufficiently small window of time to enable collaborative perception.

### 3.2 Feature Extraction

Feature extraction is a crucial step in Vehicle-to-Infrastructure (V2I) collaborative perception. After the completion of metadata sharing, the system needs to perform feature extraction on the point cloud data from each agent to generate high-dimensional feature maps for subsequent processing. To achieve real-time feature extraction while ensuring perception accuracy, this paper adopts the efficient PointPillar(Lang et al., 2019) network for feature extraction on both the vehicle and infrastructure sides. PointPillar enables the processing of large-scale point cloud data by voxelizing the point cloud into stacked pillar tensors and generating 2D pseudo-images, effectively reducing computational overhead and memory usage. The detailed process begins by transforming the raw point cloud data into a structured representation. Let $P_j^t$ denote the raw point cloud captured by agent $j$ at time $t$, which consists of a set of 3D points $\{p_1, p_2, \ldots, p_n\}$, where each point $p_i$ is described by

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-673-2025 | © Author(s) 2025. CC BY 4.0 License.

its coordinates $(x_i, y_i, z_i)$ and associated features (such as reflectance or intensity). The PointPillar network first voxelizes this point cloud data by discretizing the 3D space into a grid of pillars. Each pillar can be described as:

$$V_{x,y} = \{p_i \mid p_i \in \text{Pillar}(x, y)\} \tag{2}$$

where $\text{Pillar}(x, y)$ represents the pillar located at grid cell $(x, y)$ in the horizontal plane. For each pillar, the point features are aggregated into a fixed-size tensor through operations like mean pooling or max pooling, forming a feature tensor $T_{x,y}$. These pillar features are then fed into a 2D convolutional backbone, transforming the voxelized representation into a 2D pseudo-image:

$$I_j^t \in \mathbb{R}^{H \times W \times C} \tag{3}$$

where $H$ and $W$ are the height and width of the pseudo-image, and $C$ is the number of feature channels. This pseudo-image representation allows the system to apply standard 2D convolutional operations to the point cloud data, significantly speeding up the feature extraction process. The backbone network used in this process is a series of 2D convolutional layers that extract increasingly abstract features from the input pseudo-image. Let $\mathcal{F}(\cdot)$ denote the backbone network, and the resulting feature map for agent $j$ at time $t$ is:

$$F_j^t = \mathcal{F}(I_j^t) \in \mathbb{R}^{H' \times W' \times C'} \tag{4}$$

where $H'$, $W'$, and $C'$ represent the height, width, and number of channels of the output feature map, respectively. The feature map $F_j^t$ serves as the basis for further operations such as map-based compression and fusion with other agents' feature maps. To ensure the efficiency and scalability of the V2I collaborative perception system, the PointPillar network was chosen due to its low inference latency and high memory efficiency, as well as its robustness in handling large-scale point cloud data from both vehicle and infrastructure sources. By voxelizing the point cloud into pillars and reducing the problem to 2D feature extraction, the system can efficiently handle the large amounts of data generated in real-world V2I scenarios. In summary, the feature extraction process transforms raw point cloud data into high-dimensional feature maps through voxelization, pseudo-image generation, and 2D convolutional processing. The resulting feature maps $F_j^t$ for each agent are then used in subsequent stages of the V2I collaborative perception framework, including differential map compression and affine transformation-based decompression.

### 3.3 Communication Compression

To achieve efficient feature transmission in Vehicle-to-Infrastructure (V2I) systems, we propose a communication compression mechanism designed to filter out irrelevant background information and focus on key areas of interest, particularly dynamic objects. This mechanism aligns the map point cloud with the infrastructure point cloud by converting both into a unified spatial coordinate system. Since the map point cloud remains geographically fixed relative to the infrastructure, this transformation ensures an accurate comparison between the two point clouds.

Due to differences in density and observation angles between the map and infrastructure point clouds, as illustrated in Fig.2, we first downsample the map data and apply a larger voxelization scale to extract map voxel features, denoted as $\mathrm{F}_m = \{f_{m,j}\}_{j=1,\ldots,N}$. A trainable multi-layer compression network is then employed to compress both the map voxel features $f_{m,j}$ and the infrastructure features $f_{i,j}^t \in \mathbb{R}^{C,K,K}$. Pointwise convolution layers are used to reduce the dimensionality of these features, transforming them into query feature maps $M_{m,j}$ and $M_{i,j}^t \in \mathbb{R}^{\frac{C}{128}, \frac{K}{4}, \frac{K}{4}}$. The process is supervised by a downstream loss function, ensuring effective feature transformation. During inference, the precomputed map query feature maps $M_{m,j}$ are directly used to accelerate the process, eliminating the need for real-time point cloud transformation.

One key innovation in this method is the use of a map-mask to selectively compress the feature maps. This approach significantly reduces bandwidth while preserving essential perceptual information. The infrastructure feature map $M_{i,j}^t$ is filtered using the mask $Mask_i^t$, which focuses on extracting dynamic areas relevant to the task:

$$F_i^{\text{masked}} = \Lambda^v(M_{i,j}^t, \text{Mask}_i^t) \tag{5}$$

Similarly, the same mask is applied to the map feature map $M_{m,j}$, ensuring alignment with the infrastructure feature map:

$$F_m^{\text{masked}} = \Lambda^m(M_{m,j}, \text{Mask}_i^t) \tag{6}$$

This process guarantees consistency between the extracted regions, allowing for accurate comparison between the map and infrastructure feature maps. Once the masked feature maps are obtained, the system computes the differential feature map, capturing the significant changes between the infrastructure and map:

$$M_{i \leftrightarrow m,j}^t = \Psi(F_i^{\text{masked}}, F_m^{\text{masked}}, \lambda) \tag{7}$$

Here, $\Psi(\cdot)$ calculates the difference between the masked feature maps, while a threshold $\lambda$ filters out insignificant variations, retaining only the critical information necessary for transmission.

This multi-step process—from mask generation to differential feature extraction—enables the system to focus solely on essential dynamic information, minimizing data size and reducing communication overhead. Before the differential operation, pointwise convolution is applied to reduce the number of channels. This serves two key purposes: reducing bandwidth consumption and facilitating subsequent processing. The compressed feature maps generated through this method allow for efficient data transmission without the need to preserve all perceptual details at this stage. Instead, channel reduction optimizes the transmission of relevant data, making it easier to handle in later stages.

By leveraging this approach, the communication compression mechanism effectively minimizes bandwidth usage and prepares the system for the decompression phase. When combined with the decompression module, the system balances bandwidth efficiency with perception accuracy. The extracted dynamic object information is transmitted efficiently, supporting real-time collaborative perception with minimal bandwidth consumption while ensuring the integrity of critical information during reconstruction.

### 3.4 Communication Decompression

In the process of collaborative perception, the decompression phase plays a vital role in reconstructing compressed data received from the infrastructure. The Communication Decompression mechanism ensures that the transmitted feature maps,

Table 1. Comparison to different fusion methods on the V2X-Seq-SPD dataset, Mapcooper consistently outperforms all other fusion approaches

| Model | Fusion | mAP@3D ↑ | | mAP@BEV ↑ | | AB (Average Byte) ↓ |
|---|---|---|---|---|---|---|
| | | IoU=0.5 | IoU=0.7 | IoU=0.5 | IoU=0.7 | |
| PointPillars | VehOnly | 48.25 | 26.55 | 51.94 | 47.80 | – |
| Early Fusion | Early | 54.63 | 32.23 | 61.08 | 50.06 | $1.06 \times 10^6$ |
| Late Fusion | Late | 52.43 | 31.54 | 58.10 | 49.25 | $\mathbf{2.88 \times 10^2}$ |
| FFNet | Middle | 55.81 | 30.23 | 63.54 | 54.16 | $1.24 \times 10^5$ |
| Mapcooper | Middle | **58.90** | **33.11** | **67.08** | **59.63** | $1.02 \times 10^4$ |

which have undergone differential map compression, are accurately restored. This step leverages affine transformation matrices to map the compressed feature data back into a high-dimensional feature space. Furthermore, the map background information is integrated during decompression to provide additional context and improve the accuracy of the perception system.

The primary goal of this phase is to recover the transmitted map-mask feature map $M^t_{i\leftrightarrow m,j}$ and reintegrate it with local perception data. First, we compute affine transformation matrices $T_{i\to v} \in \mathbb{R}^{N,2,3}$, where $N$ denotes the batch size. These matrices are derived using geometric calibration parameters between the infrastructure and the vehicle. With these, we generate a sampling grid $G_{i\to v} \in \mathbb{R}^{N,K,K,2}$, representing the flow field from the infrastructure's coordinate system to the vehicle's.

The next step applies affine transformations to the feature map $M^t_{i\leftrightarrow m,j}$, allowing for the recovery of the dynamic object information. This transformation is described by:

$$M^{t,m}_{i\to v,j} = \Gamma\left(\Phi\left(M^t_{i\leftrightarrow m,j}, G_{i\to v}\right)\right) \tag{8}$$

Here, $\Phi(\cdot)$ denotes the grid sampling function, which remaps the feature map based on the transformation grid $G_{i\to v}$. The function $\Gamma(\cdot)$ represents bilinear interpolation, applied to enhance the finer details of dynamic objects after the transformation.

To reconstruct the complete feature map, we perform a similar transformation for the map-side features:

$$M_{m\to v,j} = \Gamma\left(\Phi\left(M_{m,j}, G_{m\to v}\right)\right) \tag{9}$$

This equation similarly maps the query features from the map side into the vehicle's coordinate system, ensuring alignment across different viewpoints.

Once the transformation is complete, the feature maps $M^{t,m}_{i\to v,j}$ and $M_{m\to v,j}$ are aggregated. This step is crucial for integrating the complementary information from both sources:

$$M^t_{i\to v,j} = M^{t,m}_{i\to v,j} \oplus M_{m\to v,j} \tag{10}$$

The operator $\oplus$ denotes the element-wise aggregation, which combines the information from both transformed feature maps to form the restored query feature map.

After this aggregation, the decompression process utilizes a trainable multi-layer upsampling network. This network decompresses the aggregated feature map, generating infrastructure-side features $f^t_{i\to v,j}$ that are in the same spatial coordinate system as the vehicle's local features $f^t_{v,j}$. The upsampling network ensures that the spatial and dynamic consistency is preserved:

$$F^t_{i\to v,j} = \text{Upsample}(M^t_{i\to v,j}) \tag{11}$$

This allows the infrastructure and vehicle feature maps to be fully aligned and prepared for further processing.

The final restored feature map, now enriched with both infrastructure and vehicle information, provides a comprehensive understanding of the environment. This combined feature map is then passed through a 3D object detection network, such as a Single Shot Detector (SSD), which generates the final 3D bounding boxes and classification results, including object positions, orientations, and types.

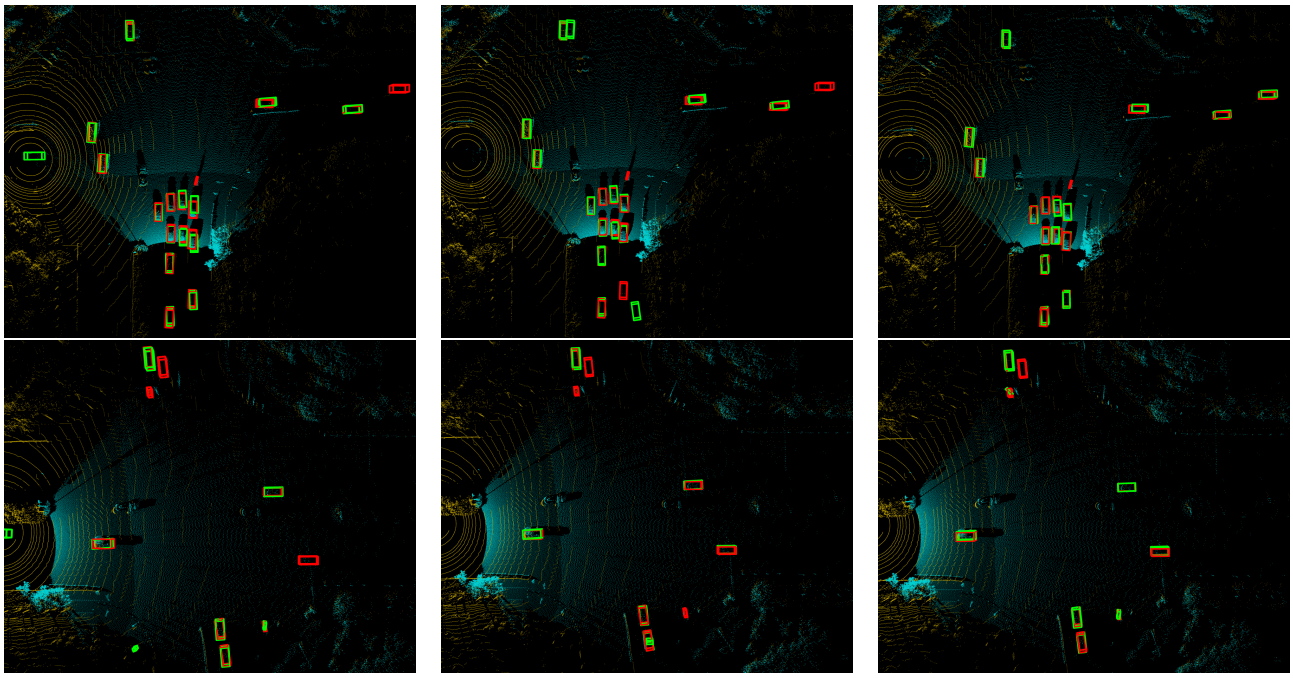
The Communication Decompression mechanism is essential for reconstructing accurate perception data. By integrating affine transformation and map background information, this process not onq ly restores the original compressed data but also enhances the perception accuracy. The enriched feature maps ensure that critical dynamic information is preserved, allowing the collaborative perception system to function effectively with minimal loss during transmission.

### 3.5 Feature Fusion and 3D Detection Head

After the decompressed critical region feature map $F^t_{i\to v}$ has been restored, the next step is to concatenate the vehicle-side feature map with the infrastructure-side decompressed feature map to achieve feature fusion for collaborative perception. Let the vehicle-side feature map be denoted as $F^t_v$. The final fused feature map can be represented as:

$$F^t_{\text{fusion}} = \text{concat}(F^t_v, F^t_{i\to v}) \tag{12}$$

Here, the *concat* operation denotes the concatenation of feature maps along the channel dimension, thus forming a unified feature representation. Through this fusion approach, the system can simultaneously leverage both vehicle-side and infrastructure-side perception information, ensuring that the perception results for the critical regions incorporate both the vehicle's local view and the infrastructure's global sensing capabilities. The fused feature map $F^t_{\text{fusion}}$ is then fed into the 3D object detection head for object detection and classification. In this work, we employ a 3D detection head based on the Single Shot Detector (SSD) architecture to perform this task. The 3D SSD head operates by applying a convolutional neural network (CNN) to process the fused features, generating 3D bounding boxes and object class predictions. The detection head maps the feature grid to each voxel in the spatial space and predicts whether an object exists in that voxel, along with the object's class and precise location in 3D space. The output for each detected object $i$ includes both its 3D bounding box and its classification result. The detection for the $i$-th object can be



(a) veh only (b) FFNet (c) Mapcooper

Figure 3. The detection visualization results of Veh Only, FFNet and Mapcooper across several challenging scenarios.

expressed as:

$$\text{BBox}_i = \{(x_i, y_i, z_i), (w_i, h_i, d_i)\} \tag{13}$$

$$\text{Class}_i = \arg\max(\text{softmax}(c_i)) \tag{14}$$

where $(x_i, y_i, z_i)$ denotes the 3D coordinates of the object's center, $(w_i, h_i, d_i)$ represent the width, height, and depth of the object along the $x$, $y$, and $z$ axes, respectively, and $c_i$ is the class score vector. The object class $\text{Class}_i$ is determined as the class with the highest score from the softmax function. Through the above process, the system finally outputs a set of 3D bounding boxes and corresponding object classification results. This information provides the vehicle with precise location data of potential obstacles or objects in its surrounding environment, aiding in further path planning and decision-making.

## 4. Experiments

### 4.1 Dataset

**V2X-Seq-SPD**: The V2X-Seq-SPD(Yu et al., 2023) dataset is the first real-world dataset designed for V2X sequential perception, capturing cooperative data between vehicles and infrastructure. It consists of approximately 100 sequences of images and point cloud data from both perspectives, with comprehensive annotations including 3D bounding boxes and tracking IDs. In total, the dataset spans over 15,000 frames from 95 distinct temporal sequences. This paper focuses primarily on 3D object detection using the V2X-Seq-SPD dataset. Additionally, the temporal characteristics of the dataset were leveraged by applying the Direct LiDAR Odometry(Chen et al., 2022) algorithm to map six intersections, providing valuable data for experimental purposes.

### 4.2 Experimental Setup

**Implementation details**: We implemented our experiments using MMDetection3D as the primary framework, training the baseline model on the V2X-Seq-SPD dataset for 40 epochs. The initial learning rate was set to 0.001, and a weight decay of 0.01 was applied for optimization. Model training and evaluations were conducted on an NVIDIA GeForce RTX 4090 GPU. During the experiments, detection analysis was performed solely on dynamic objects located within the predefined rectangular area $[0, -39.12, 100, 39.12]$.

**Evaluation metrics**

*Perception Performance*: For 3D object detection evaluation, we adopted the KITTI benchmark standards, including $AP_{3D}$ (3D Average Precision) and $AP_{BEV}$ (Bird's Eye View Average Precision). Average Precision (AP) was used to assess detection performance at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7. For this evaluation, vehicles detected by at least one connected LiDAR were considered.

*Communication Bandwidth*: We employed $AB$ (average Byte) as the metric to assess transmission costs, disregarding calibration files and timestamps. The overall transmission cost was evaluated based on the transmission of raw data, detection outputs, or feature tensors, with the transmission cost calculated per frame to quantify the bandwidth consumption.

### 4.3 Quantitative Evaluation

The experimental results highlight the significant advantages of Mapcooper in both detection accuracy and communication efficiency. As shown in Table 1, Mapcooper surpasses non-fusion methods like PointPillars, with a notable 10.65% improvement in mAP@3D (IoU=0.5) and an 15.14% increase in mAP@BEV (IoU=0.5). This demonstrates the benefit of incorporating infrastructure data into the perception process, improving the system's ability to detect objects in its surroundings. Although late fusion models are effective in reducing transmission costs, they compromise on perception performance. In contrast, Mapcooper outperforms late fusion by 6.47% in mAP@3D (IoU=0.5) and 8.98% in mAP@BEV (IoU=0.5),

underscoring that communication efficiency alone is insufficient if it results in reduced detection accuracy. Compared to early fusion methods, Mapcooper achieves superior detection accuracy in both mAP@BEV and mAP@3D while dramatically reducing bandwidth requirements to only 1% of what early fusion consumes. This shows Mapcooper's ability to balance high performance with lower communication overhead, making it highly suitable for bandwidth-limited environments. Finally, compared to other middle fusion approaches like FFNet, Mapcooper delivers state-of-the-art detection results, with a 3.09% improvement in mAP@3D (IoU=0.5) and a 3.54% increase in mAP@BEV (IoU=0.5), while using only 1/12th of the transmission cost. These results firmly establish Mapcooper's efficiency in achieving both optimal perception performance and minimal communication cost, making it an ideal solution for systems requiring high detection accuracy and bandwidth efficiency.

### 4.4 Qualitative evaluation

**Detection visualization**:Fig.3 shows the detection visualizations of Veh Only, FFNet, and Mapcooper across several challenging scenarios. Mapcooper consistently demonstrates enhanced performance in identifying distant and smaller objects within sparse point clouds. For instance, in the first scenarios, only Mapcooper successfully detects the vehicle on the far-right side, which is missed by the other models. Additionally, in the second scenario, Mapcooper accurately identifies smaller objects that are undetected by other methods. This improvement is largely attributed to the integration of the Map Mask module, which effectively enriches the contextual information around dynamic objects. By incorporating additional background context, Mapcooper is better equipped to detect objects that are otherwise hard to perceive, particularly in scenarios with sparse or non-salient targets. These results highlight Mapcooper's superior detection accuracy compared to alternative approaches, especially in situations where capturing subtle features is essential.

## 5. Conclusion

In this paper, we presented Mapcooper, a unified framework for V2I collaborative perception designed to optimize the balance between communication bandwidth and perception performance. The proposed method leverages a novel map alignment communication strategy (MACS) to filter out irrelevant background information during the compression phase, ensuring that only critical perceptual features are transmitted. This approach significantly reduces communication overhead while maintaining high perceptual accuracy by restoring the contextual scene information in the decompression phase. Through extensive experimentation on the large-scale V2X-Seq-SPD dataset, we demonstrated that Mapcooper enhances perception performance with an average accuracy improvement of 4.59% over existing state-of-the-art methods, while simultaneously reducing bandwidth consumption to just 1/12 of that used by the most bandwidth-efficient prior approaches. These results highlight the potential of Mapcooper in real-world applications, offering a scalable and efficient solution for intelligent transportation systems that demand high situational awareness and resource efficiency. Future work will explore further enhancements to the map-mask filtering mechanism, focusing on dynamic adaptability to varying environmental conditions and communication constraints.

## References

Chen, K., Lopez, B. T., Agha-mohammadi, A.-a., Mehta, A., 2022. Direct lidar odometry: Fast localization with dense point clouds. *IEEE Robotics and Automation Letters*, 7(2), 2000–2007.

Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S., 2019a. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 88–100.

Chen, Q., Tang, S., Yang, Q., Fu, S., 2019b. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 514–524.

Cheng, J., Yin, W., Wang, K., Chen, X., Wang, S., Yang, X., 2024. Adaptive fusion of single-view and multi-view depth for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10138–10147.

Chuah, W., Tennakoon, R., Hoseinnezhad, R., Suter, D., Bab-Hadiashar, A., 2022. Semantic guided long range stereo depth estimation for safer autonomous vehicle applications. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 18916–18926.

Fang, J., Zhou, D., Song, X., Zhang, L., 2021. Mapfusion: A general framework for 3d object detection with hdmaps. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 3406–3413.

Fujimoto, T., Tanaka, S., Kato, S., 2022. Lanefusion: 3d object detection with rasterized lane map. *2022 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 396–403.

Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S., 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35, 4874–4886.

Huang, Y., Zhou, J., Li, X., Dong, Z., Xiao, J., Wang, S., Zhang, H., 2023. MENet: Map-enhanced 3D object detection in bird's-eye view for LiDAR point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 120, 103337.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O., 2019. Pointpillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W., 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34, 29541–29552.

Liu, Y.-C., Tian, J., Glaser, N., Kira, Z., 2020. When2com: Multi-agent perception via communication graph grouping. *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 4106–4115.

Rawashdeh, Z. Y., Wang, Z., 2018. Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 3961–3966.

Wang, T., Chen, G., Chen, K., Liu, Z., Zhang, B., Knoll, A., Jiang, C., 2023. Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8187–8196.

Wang, T.-H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R., 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 605–621.

Xiao, J., Wang, S., Zhou, J., Tian, Z., Zhang, H., Wang, Y.-F., 2024. MIM: High-Definition Maps Incorporated Multi-View 3D Object Detection. *IEEE Transactions on Intelligent Transportation Systems*.

Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J., 2022a. Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*.

Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.-H., Ma, J., 2022b. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. *European conference on computer vision*, Springer, 107–124.

Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z., 2022. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. *European Conference on Computer Vision*, Springer, 677–695.

Yang, B., Liang, M., Urtasun, R., 2018. Hdnet: Exploiting hd maps for 3d object detection. *Conference on Robot Learning*, PMLR, 146–155.

Yin, T., Zhou, X., Krahenbuhl, P., 2021. Center-based 3d object detection and tracking. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11784–11793.

Yu, H., Tang, Y., Xie, E., Mao, J., Luo, P., Nie, Z., 2024. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems*, 36.

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

Zeng, Z., Qiu, H., Zhou, J., Dong, Z., Xiao, J., Li, B., 2024. PointNAT: Large Scale Point Cloud Semantic Segmentation via Neighbor Aggregation with Transformer. *IEEE Transactions on Geoscience and Remote Sensing*.