SCORPION: Robust Spatial-Temporal Collaborative Perception Model on Lossy Wireless Network

Ruiyang Zhu[1], Minkyoung Cho[1], Shuqing Zeng[2], Fan Bai[2], Z. Morley Mao[1]

***Abstract*— Collaborative Perception enables multiple agents, such as autonomous vehicles and infrastructure, to share sensor data via vehicular networks so that each agent gains an extended sensing range and better perception quality. Despite its promising benefits, realizing the full potential of such systems faces significant challenges due to inherent imperfections in underlying system layers, consisting of network layer imperfections and hardware-level noises. Such imperfections and noises include packet loss in vehicular networks, localization errors from GPS measurements, and synchronization errors caused by clock deviation and network latency. To address these challenges, we propose a novel end-to-end collaborative perception framework, SCORPION, that harnesses the AI co-design of the application layer and system layer to tackle the aforementioned imperfections. SCORPION consists of three main components: lost bird's eye view feature reconstruction (L-BEV-R) recovers lost spatial features during lossy V2X communication, while deformable spatial cross attention (DSCA) and temporal alignment (TA) compensate for localization and synchronization errors in feature fusion. Experimental results on both synthetic and real-world collaborative 3D object detection datasets demonstrate that SCORPION advances the state-of-the-art collaborative perception methods by 5.9 - 13.2 absolute AP on both standard and noisy scenarios.**

I. Introduction

Connected and autonomous vehicles (CAV) rely heavily on their onboard sensors to perceive and navigate in complex driving environments. However, single-vehicle perception systems often struggle with occlusions and limited sensing ranges, potentially leading to severe consequences [3]. Collaborative Perception [4], [5], [6], [7], [8], [9], [2], [10], [11], [12] addresses the limitation of single-agent sensing by allowing multiple agents (vehicles/infrastructures) to share sensor data via vehicular networks so that each vehicle gains an extended sensing range and a better understanding of spatial semantics, particularly in occluded areas. Recent research also highlights that incorporating temporal consistency enhances the perception of dynamically moving objects [6], [13], [14]. Thus, robust and accurate perception relies on comprehending both spatial and temporal information for a complete contextual understanding.

However, it's important to note that collaborative perception relies on networks and hardware devices that are subject to real-world constraints, significantly impacting both temporal and spatial information. Realizing collaborative perception thus requires a deep understanding of these underlying systems. For instance, packet drops in Vehicle-to-Everything (V2X) communication result in lost spatial semantics, leading to missing crucial information needed for decoding the final perception output, while network delays can cause temporal misalignment during data fusion. Similarly, hardware devices such as GPS+RTK [15] and onboard clock devices often produce noisy measurements: localization errors in vehicle pose can lead to spatial misalignment [16], and clock deviation [6] can introduce temporal misalignment. Fig. 1 and 2 illustrate the impact of these issues, highlighting the need for AI solutions that account for the fundamental limitations of underlying stacks.

Historically, efforts to enhance network reliability [17], [18] and vision application performance [19], [20] were often pursued independently, reflecting a longstanding division between network and application-layer advancements. Previous work on collaborative perception application (Table I) often focus on improving perception performance by designing advanced fusion mechanisms [7], [1] and strategies to reduce network bandwidth usage [4]. All these solutions assume that the communication and perception environments are ideal. A few recent work started to look at real-world noises such as pose noise (localization noise) [2], [5], [13], synchronization noise [6], and lossy communication [21]. However, *none* of these approaches take a holistic view of the entire system stack, including fundamental constraints such as packet encoding in vehicular networks or the interplay between hardware noise and network impairments. This gap significantly limits the practical deployment and scalability of collaborative perception in real-world applications.

To bridge the gap, in this work, we present SCORPION , a robust **S**patial-temporal **CO**llabo**R**ative **P**erceptI**o**n model on l**O**ssy **N**etwork. SCORPION is an end-to-end approach that embodies the nascent philosophy of AI co-design across the application and lower system layers (including lossy network and noises of hardware devices) to directly address and compensate for the inherent imperfections from these system layers. At the core of SCORPION is a bird's-eye view (BEV)-based intermediate fusion: BEV effectively captures the spatial layout with a consistent top-down view, facilitating effective fusion from different viewpoints; meanwhile, intermediate fusion strikes a balance between network resource demands and accuracy, making it a practical approach in collaborative perception [7], [2]. Based on the BEV feature map, our three main components enhance spatial and temporal understanding: (1) For spatial semantics, our lost BEV feature reconstruction module incorporates realistic network packet encoding and recovering strategies that can recover spatial semantics lost during wireless V2X communication;

[1]University of Michigan, Ann Arbor, USA {ryanzhu, minkycho, zmao}@umich.edu
[2]General Motors LLC, Warren, MI, USA {shuqing.zeng, fan.bai}@gm.com

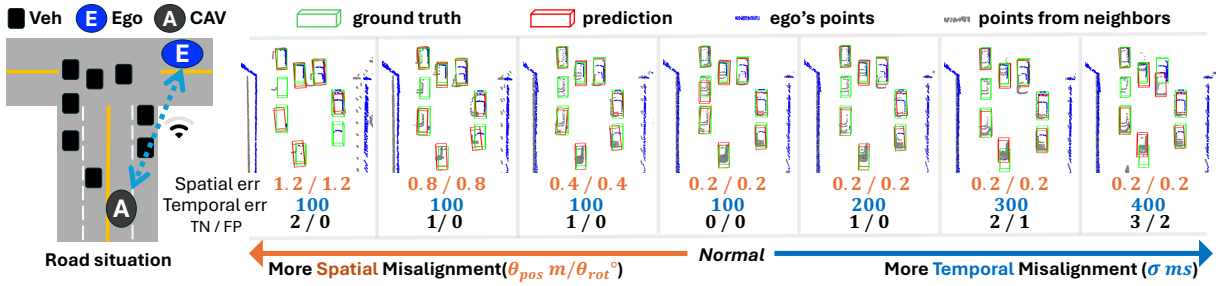

Fig. 1. Motivating examples of various degrees of spatial and temporal misalignments. Spatial misalignments occur due to sensing errors ($\star$) or dropped network packets, while temporal misalignments arise from sensor asynchronization ($\star$) and network delays. For simplicity and clarity, we focus on the cases marked with ($\star$) in this figure. This scenario is generated using the CoBEVT [1] model on the V2XSet Dataset [2]. TN and FP represent the True Negative and False Positive detections made by the model.

(2) for temporal consistency, temporal alignment module align synchronization errors by leveraging historical frame information; and (3) for the correlation between temporal features and spatial features, a deformable spatial cross-attention module facilitates flexible mapping between ego vehicle's historical and neighbor's features. By integrating these components into a single transformer [22] architecture, SCORPION demonstrates notable improvements over state-of-the-art (SOTA) methods on the OPV2V, V2XSet, DAIR-V2X datasets [7], [2], [23] in both standard and challenging scenarios, achieving robust collaborative perception.

The main contributions of this work are as follows:

- Real-world networking and hardware challenges, like lossy communication and sensor noise, are typically addressed within domain-specific boundaries [6], [24], [5]. SCORPION demonstrates the new potentials of transformer-based AI to tackle these persistent issues, expanding its scope to enable co-design across the entire hardware-network-AI software stack.
- SCORPION achieves robust collaborative perception by introducing a novel feature masking and reconstruction strategy specifically designed for vehicular networks (lost feature reconstruction), integrated with a deformable spatial cross-attention module and a temporal alignment module, allowing simultaneous tolerance to dynamic noisy environments with lossy V2X communication, localization errors, and synchronization errors.
- SCORPION outperforms SOTA methods in both standard and challenging scenarios, including extreme sensing errors and congested network conditions. For example, on the V2XSet dataset [2], SCORPION achieves 13.2%, 7.5%, and 5.9% higher accuracy than the SOTA approaches in situations involving lossy networks, spatial misalignment, and temporal misalignment.

II. Background and Related Work

Collaborative Perception. Collaborative perception solutions can be divided into three categories based on message-sharing strategy: (1) Early fusion, (2) Intermediate fusion, and (3) Late fusion. Late fusion integrates decisions from independent models at the object level [26] and tends to yield suboptimal accuracy due to the lack of cross-sensor feature enrichment [27]. Early fusion shares raw sensor data, which is network-resource intensive. To reduce bandwidth cost, existing methods [24], [6], [9] use partitioning, where data from each vehicle's nearby region is transmitted and merged to create an extended view of the environment.

TABLE I

Comparison of Different Works Handling System Imperfections.

Work	Sensing Errors	Sensor Asynchronization	Lossy V2X Network	Fusion
EMP [24]	✗	✗	✗	Early
Harbor [9]	✗	✗	✗	Early
Where2comm [4]	✗	✗	✗	Intermediate
CoBEVT [1]	✗	✗	✗	Intermediate
RAO [6]	✗	✓	✗	Early
V2X-ViT [2]	✓	✓	✗	Intermediate
CoAlign [5]	✓	✗	✗	Intermediate
LCRN [21]	✗	✗	✓	Intermediate
SCOPE [13]	✓	✓	✗	Intermediate
VIPS [25]	✗	✓	✗	Late
SCORPION (Ours)	✓	✓	✓	Intermediate

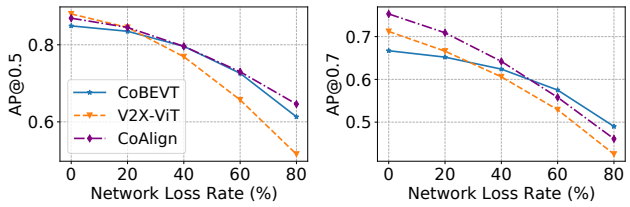

Fig. 2. Illustration of detection performance drop suffering from lossy V2X wireless communication based on the V2XSet dataset [2]. The three baseline methods, CoBEVT [1], V2X-ViT [2], and CoAlign [5] are trained with no network packet loss and tested at various level of loss.

Intermediate fusion shares encoded features, reducing the network resources needed compared to raw data. For effective feature fusion, AttFuse [7] applies self-attention to analyze features at the same spatial locations across agents. CoBEVT [1] shares Bird's Eye View (BEV) features to efficiently represent multi-view data and uses transformers to fuse BEV features. Where2Comm [4] optimizes bandwidth by identifying and sharing only essential information. Beyond accuracy and bandwidth efficiency, some works propose robust models to improve resilience in error-prone scenarios. V2X-ViT [2] introduces an advanced attention model to handle asynchronous error in V2X communication. LCRN [21] achieves robust vehicle-to-vehicle communication under lossy communication by employing a repair network. CoAlign [5] addresses pose estimation errors through agent-object pose graph modeling to ensure pose consistency across agents. While temporal and spatial misalignment and lossy communication are common issues in reality, no current work addresses all of these challenges comprehensively

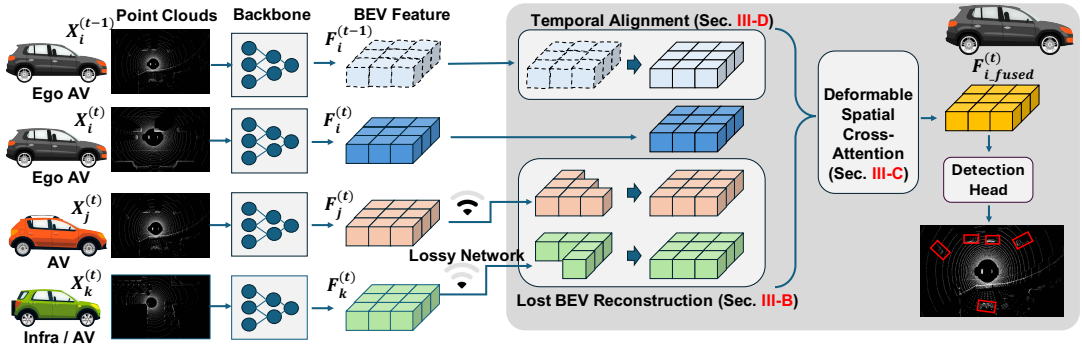

Fig. 3. The overall architecture of SCORPION. The framework consists of three major components: a lost BEV feature reconstruction (L-BEV-R) module, a deformable spatial cross-attention (DSCA) module, and a temporal alignment (TA) module.

(see Table I). This motivates us to develop a more robust collaborative perception network for practical scenarios.

Vehicle-to-Everything (V2X) Network. Collaborative perception relies on transmitting sensing information between agents. Technologies like DSRC [28] and C-V2X [29] are designed for V2X communication. However, these wireless networks are inherently unreliable. Communication on vehicular networks often relies on UDP [30] as the underlying transport protocol [31], [32], [33], which can have a high loss rate under dynamic and fast-changing vehicle topologies. Prior work on collaborative perception [2], [7], [4], [1], [5], [6], [11] assumes that the transmission of feature map or raw data will not get lost, which is not realistic on a real-world V2V/V2X setup [33], [32]. The only prior work explicitly addressing lossy communication [21] oversimplifies the issue by simulating network loss with random noise while *ignoring the structured packet encoding of shared feature maps*, making it impractical for real deployment. Therefore, we argue that tolerance to practical lossy V2X communication is a vital characteristic that collaborative perception models must address, as incomplete shared intermediate features from lossy communication can significantly compromise effectiveness (Fig. 2).

III. Method

In this paper, we focus on enhancing the robustness and accuracy of collaborative perception, emphasizing temporal and spatial understanding in the face of inherent system imperfections. We propose a novel BEV-based intermediate fusion framework SCORPION. Our framework consists of three key components (Fig. 3): Lost BEV Feature Reconstruction (L-BEV-R), Deformable Spatial Cross-Attention (DSCA), and Temporal Alignment (TA).

A. Problem Formulation

At time t, N agents participate in a collaborative perception task, establishing a communication graph. We designate the i-th agents as the ego vehicle, with other agents within the communication range acting as collaborators. These collaborators share perceptual information and metadata (*e.g.,*pose, timestamps) with the ego vehicle for collaboration. Each agent $j \in \{1, ..., N\}$ encodes the point clouds into Bird's Eye View (BEV) features to extract local visual features. Given the point cloud $X_j^{(t)}$ of j-th agent at timestamp t,

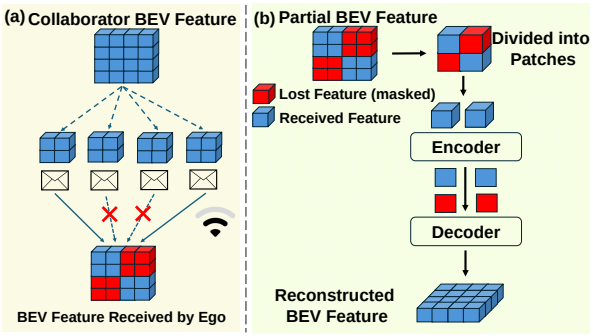

Fig. 4. Our Lost BEV Feature Reconstruction architecture.

the extracted feature is $F_j^{(t)} = f_{enc}(X_j^{(t)}) \in \mathbb{R}^{C \times H \times W}$, where $f_{enc}(\cdot)$ is the feature encoder shared among all agents, and C, H, W stand for the channel, height, and width. The collaborator agents share the processed feature maps with the ego agent for fusion and perception. In a realistic setup, feature maps go through lossy V2X communication, and poses or timestamps of agents can contain noise. We show how SCORPION handles the lost spatial features in §III-B, how SCORPION addresses the misaligned spatial information in §III-C, and how to maintain temporal consistency in §III-D.

B. Lost BEV Feature Reconstruction

Reconstructing the original signal from partial observation of images has been a challenging task in computer vision. In the context of realistic collaborative perception, our goal is to address the critical issue of data loss in V2X communications, a gap not effectively bridged by current methods [2], [1], [5]. Inspired by the masked-autoencoder (MAE) [34], we design a customized Lost-BEV-Reconstruction (L-BEV-R) network for intermediate feature recovery from other CAVs. We treat the intermediate BEV feature map after the lossy communication as partial images, where the locations with features that are received as *visible patches* and treat the locations of the features that get lost during V2X network transmission as *masked patches* that require recovering.

The framework of the L-BEV-R network is shown in Fig. 4, which is an encoder-decoder architecture based on Vision Transformers [35]. We next elaborate on the details.

V2X Network Packet Encoding and Partitioning for Feature Maps. Previous work [36] on the system and network community has proposed methods to encode raw point cloud data into self-contained network packets. To the best of our

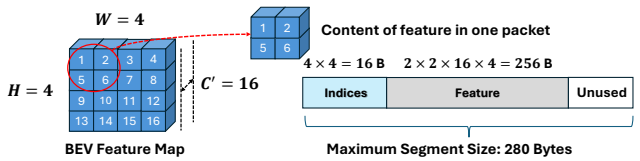

Fig. 5. An example of an encoded feature map into a network packet.

knowledge, no prior work has discussed how to encode and partition the intermediate feature maps into network packets. In SCORPION, we detail the process of packing the Bird's-Eye-View (BEV) features into multiple *self-contained network packets*. In the context of intermediate-fusion collaborative perception, network packet loss impacts by dropping feature map values. Assuming a BEV feature map of size $\mathbb{R}^{C\times H\times W}$, the actual transmission of the feature will break the feature map into multiple network packets (Fig. 4(a)), where the size of each packet is limited by Maximum Segment Size (MSS). For V2X communication, MSS is in the range of 50 bytes to 300 bytes [37], [36]. For example, assuming a feature map of shape $(C', H, W) = (16, 4, 4)$ where C' is the channel size after compression [7] and an MSS of 280 bytes (Fig 5), a packet can hold information on $\lfloor 280/((16+1)\times 4)\rfloor = 4$ cells in the BEV map, where "+1" accounts for additional metadata of feature indices.

Fundamental Similarity of Network Packet Loss and Feature Masking. As a network packet can cover a small range of geometric locations in the BEV feature map, we can see that packet losses are fundamentally similar to masking input images in the MAE architecture. In this case, a packet loss during the V2X communication indicates that the ego vehicle will lose the 4 cells in the received feature map. Based on this, our masking strategy in the L-BEV-R module is straightforward: we place patches on the feature maps, treat the patches that contain lost indices as masked patches, and train an encoder-decoder network to recover them.

L-BEV-R Encoder. Our encoder is a variation of the ViT [35] by extending the input space from 3-channel RGB images to C channel BEV feature maps. Similar to a standard ViT, we divide the feature map into patches of size $P \times P$ and add positional encoding to the patches generated from a linear projection layer. A high portion (*e.g.*,75%) of random patches will be masked during the training and the encoder operates on the visible patches through a series of Transformer blocks.

L-BEV-R Decoder. The input to the decoder is the full set of patch-tokens consisting of both encoded feature patches and masked patches representing the features that are lost during V2X communication. The decoder has another series of Transformer blocks to reconstruct a feature map of (C, H, W). The reconstruction target is the pixel value of the patch and is linearly projected to the original feature map channel C. During the training of the L-BEV-R model, we use the processed BEV feature maps generated from the backbone network [7] as the ground truth and use the mean squared error (MSE) as the loss function and compute the loss only on masked patches. Later, we fixed the parameters of the trained L-BEV-R model in the training of the entire end-to-end collaborative perception model. During testing and inference, the L-BEV-R module processes the partial BEV feature from collaborators by reconstructing the original feature map before fusing them with the ego feature.

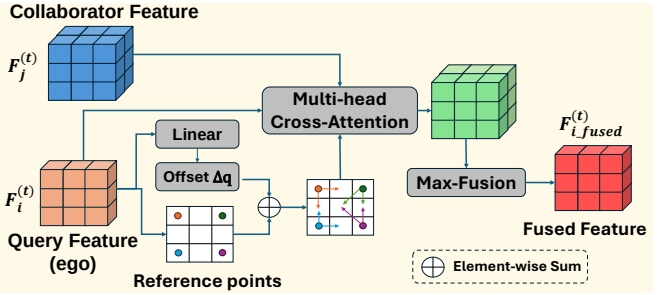

Fig. 6. The architecture of the proposed deformable spatial cross-attention (DSCA) module.

C. Deformable Spatial Cross-Attention

After reconstructing the partial feature map shared by collaborators, we proceed to fuse the multi-agent feature to enhance the perception capability over single-agent perception. Existing works [2], [7] propose to use attention-based mechanisms to fuse multi-agent feature maps and achieve SOTA performance in ideal environments. However, the assumption that feature interaction only happens at the same spatial location (*i.e.,*the same indices in the feature map in the ego agent's coordinate system) makes these methods vulnerable to localization and synchronization errors, as these errors may create misalignment of perceptual critical areas. To mitigate such effects, we draw inspiration from [38] and implement a simple yet effective Deformable Spatial Cross-Attention (DSCA) component. The DSCA module consists of deformable attention along with a max fusion aggregation of multi-agent features (Fig. 6). Different from prior work [13] that employs deformable attention *locally on only high confidence regions*, DSCA in SCORPION models the interaction of features *globally from the entire feature map* that contains localization and synchronization error and applies a max-pooling layer to aggregate most salient features, resulting in better perception performance at higher noise levels (see Sec. IV-B).

Deformable Spatial Cross-Attention Module. As depicted in Fig. 6, we treat the ego feature $F_i^{(t)}$ as the query feature Q and obtain a default reference point sets (same as the naive attention in [2]). To resolve the misalignment of BEV feature maps, the DSCA performs multi-head cross-attention on a set of sampled offset points. The *deformable* concept helps in utilizing and learning the spatially spread-out features caused by aforementioned misalignments. For each reference query point q (*i.e.,*each cell in the BEV feature grid), we select X points based on a learned offset and extract their features as the attending features. A linear layer is used to learn the 2D spatial offset Δq of the reference point q for sampling positional offset at $q + \Delta q$. The output at each query 2D position from ego feature $F_i^{(t)}$ to a collaborator feature $F_j^{(t)}$ is shown in Equation 1.

$$DSCA(q)_{i,j} = \sum_{a=1}^{A} W_a \left[\sum_{x=1}^{X} \phi(W_b F_i^{(t)}(q)) F_j^{(t)}(q + \Delta q_x) \right] \quad (1)$$

where A is the total number of attention heads and a indexes the attention head, $W_{a/b}$ are learnable weights for each ego-collaborator pair, and $\phi(\cdot)$ is the softmax function.

Max Fusion. For each pair of the ego BEV feature and a collaborator BEV feature, we generate a deformed cross-attention map at each spatial location q. To aggregate the attention features, we further apply a max-fusion at each query point to extract the most salient feature. The final output from the DSCA module at each query position is

$$DSCA(q) = \max(DSCA(q)_{i,1}, DSCA(q)_{i,2}, ..., DSCA(q)_{i,N}) \quad (2)$$

D. Temporal Alignment

Integrating historical observations has the potential to enhance the temporal consistency of collaborative perception. To capitalize on this potential, we treat past BEV feature maps as additional input sources. These historical features are aligned to the current timestamp using the ego vehicle's motion data. This alignment allows us to perform deformable cross-attention alongside the reconstructed feature maps from other agents, effectively leveraging both past and present data to minimize errors and improve the robustness of the perception system.

Historical Feature Alignment. Given the BEV queries Q at current timestamp t and a history BEV feature $F_i^{(t-1)}$ observed at time $t-1$, we first align the historical feature to the current timestamp using ego-agent motion with GPS+IMU pose to make the features at the same BEV grid correspond to the same real-world location. Note that such alignment using ego-motion handles the shifts of static environments but not dynamic objects in the scene. Inspired by [39], we model the temporal correlation between features by using a deformable temporal self-attention (DSA) layer, as follows:

$$DSA(q_p, \{F_i^{(t)}, F_i^{(t-1)}\}) = \sum_{V \in \{F_i^{(t)}, F_i^{(t-1)}\}} \text{DeformAttn}(q_p, V) \quad (3)$$

where q_p denotes the BEV query located at $p = (x, y)$ and DeformAttn is the same deformable attention module in the DSCA module. In our implementation, we merge the DSA layer with the DSCA by stacking $F_i^{(t-1)}$ with $F_j^{(t)}$, $j \in \{1, ..., N\}$ as they are conceptually same but with different parameter and meaning (*i.e.,*cross-agent attention *vs.*temporal self-attention). The fused feature map is then obtained from the max fusion of DSCA output $F_{i\_fused}^{(t)} \in \mathbb{R}^{C \times H \times W}$ and goes through a series of standard transformer layers including LayerNorm [22] and MLP [40].

E. Detection Head

We leverage both the regression decoder and classification decoder to decode the final fused feature map $F_{i\_fused}^{(t)}$ into the predicted outputs. The regression and classification decoders are MLP layers. The regression output is $\mathbb{R}^{7 \times H \times W}$, where the first dimension 7 represents the position (x, y, z), size (h, w, l), and yaw angle θ of the predicted bounding boxes at each location. The classification output is the confidence level of each predefined box to be a target vehicle or background. Following existing work [19], we adopt the smooth L1 loss for regressing the bounding boxes and a focal loss [41] for classification.

TABLE II

PERFORMANCE COMPARISON ON THE V2XSET, OPV2V, AND DAIR-V2X DATASETS IN PERFECT ENVIRONMENTS.

Model	V2XSet		OPV2V		DAIR-V2X	
	AP0.5	AP0.7	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	65.73	52.57	69.38	56.40	63.04	47.39
V2VNet [8]	87.82	74.28	86.76	73.38	65.09	48.18
F-Cooper [10]	82.82	69.38	89.22	79.66	70.54	52.21
AttFuse [7]	81.70	66.24	88.54	72.91	68.02	48.40
CoBEVT [1]	81.00	65.06	88.99	72.80	67.61	55.51
V2X-ViT [2]	82.32	71.21	86.74	75.70	70.87	54.35
CoAlign [5]	86.90	75.31	91.60	82.30	74.02	**56.81**
SCOPE [13]	87.55	75.67	89.60	80.71	74.15	56.52
SCORPION	**88.32**	**77.78**	**93.10**	**85.10**	**74.65**	56.76

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

Datasets. We validate our proposed method SCORPION with LiDAR-based 3D object detection on three datasets: two simulation datasets **V2XSet** [2], **OPV2V** [7] and a real-world vehicle-to-infrastructure dataset **DAIR-V2X** [23].

Implementation Details. We build SCORPION based on the Pytorch [42] and OpenCOOD framework [43] and train the models on NVIDIA A100 GPUs [44]. Pointpillar [19] is used as the backbone feature encoder with 0.4m $\times$ 0.4m voxel size. We train our L-BEV-R network following the standard MAE network training [34] for 100 epochs and freeze the parameters in the end-to-end model training.

Evaluation Metrics. Following existing work [2], [7], we adopt the Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 for evaluation.

Benchmark Comparison. We compare with seven state-of-the-art (SOTA) intermediate fusion models: V2VNet [8], F-Cooper [10], AttFuse [7], CoBEVT [1], V2X-ViT [2], CoAlign [5] and SCOPE [13]. In addition, we also compare our approach with single-agent perception (*i.e.,*No Fusion).

B. Quantitative Results

We compare the performance of 3D object detection with other methods (Table II). Our proposed approach SCORPION achieves the highest perception performance compared to the baseline methods on all datasets. It achieves average improvements of 22.6%/25.2%, 23.8%/28.7%, and 11.6%/9.4% on three datasets in AP@0.5/0.7 compared with 'No Fusion', showing the benefits of collaborative perception. Compared to prior SOTA collaborative perception methods, our method improves AP@0.7 by over 2.4% on the V2XSet dataset, 2.8% on the OPV2V dataset and achieves comparable performance with the best method on DAIR-V2X dataset.

Performance with Network Loss, Localization and Synchronization Noise Together. Using data from V2X network measurement [32], GPS errors [45], we evaluate SCORPION under the *coexistence of packet loss, localization, and synchronization error*. We summarized 5 diverse road settings with a variety of loss rate (12% - 32%): (1) Urban Freeway (UF), (2) Rural Freeway (RF), (3) Rural Road (RR), (4) Suburban Road (SR), (5) Open Field (OF). The network

TABLE III

PERFORMANCE COMPARISON UNDER COMBINED NETWORK LOSS, LOCALIZATION AND SYNCHRONIZATION NOISE.

		V2XSet				OPV2V				DAIR-V2X			
Case	Loss Rate	SCORPION	SCOPE	CoAlign	V2X-ViT	SCORPION	SCOPE	CoAlign	V2X-ViT	SCORPION	SCOPE	CoAlign	V2X-ViT
UF	26%	**84.0/61.2**	78.7/57.3	68.4/39.0	72.9/52.7	**88.6/76.4**	84.6/70.5	74.9/59.0	80.5/63.1	**72.5/54.3**	70.7/52.8	70.3/53.8	68.6/53.3
RF	20%	**84.2/61.9**	78.3/57.5	68.7/38.6	74.3/53.2	**88.9/77.1**	84.8/70.5	76.0/60.0	80.9/63.0	**72.9/54.5**	70.9/53.2	70.1/54.1	69.6/53.7
RR	32%	**82.0/61.0**	77.3/57.0	68.6/38.5	71.3/52.1	**88.4/75.9**	84.5/70.3	73.2/57.6	80.1/62.8	**72.1/54.3**	69.9/53.0	70.5/53.2	68.5/53.3
SR	15%	**85.2/63.1**	78.6/56.5	67.7/38.9	75.5/53.6	**89.3/77.7**	84.9/70.8	77.2/61.0	81.0/63.9	**73.2/55.2**	71.5/54.8	70.1/55.1	69.0/53.5
OF	12%	**84.8/64.9**	78.8/56.6	67.4/39.1	76.0/53.7	**89.6/78.0**	85.1/70.9	78.0/61.5	81.1/63.9	**73.5/55.5**	71.8/54.2	71.1/55.0	69.2/53.5

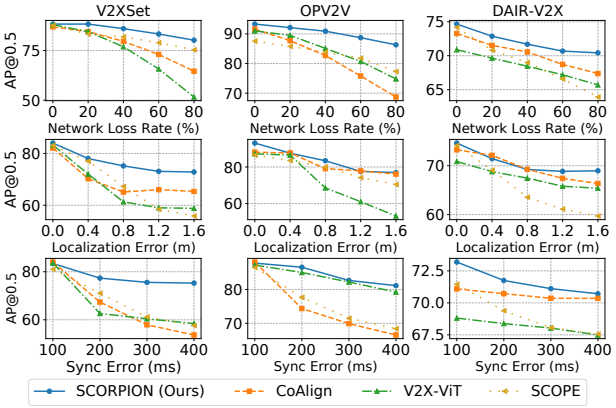

Fig. 7. Performance comparison across models under various network loss/localization/synchronization errors.

packet loss is simulated by randomly masking out the shared feature maps. We simluate localization error by adding Gaussian noise with a standard deviation of 0.6 m for urban areas and 0.4 m for rural areas and open fields following statistics from [45]. For synchronization latency, we set it to 100ms following existing work [2], [46]. We select the *top 3* performing baseline methods from the perfect environment setup, V2X-ViT, CoAlign and SCOPE for comparison in later experiments. Note that the selected baselines have partial designs on handling the system noises (Table I).

From the results in Table III, we note that SCORPION outperforms existing baselines by over 4%, 2%, and 5% for V2XSet, OPV2V, and DAIR-V2X dataset in absolute AP when **imperfect feature transmission, localization noises, and synchronization errors exist together**. These improvements stem from our holistic design, which effectively handles feature loss (via L-BEV-R) and aligns misaligned features caused by localization and synchronization errors (via DSCA and TA).

C. Robustness to Individual Noise Factor

In this section, we evaluate the robustness of SCORPION by analyzing its performance under different levels of individual noise factor: network loss, localization errors, and synchronization errors.

Results under Various Network Loss Rates. Fig. 7 reveals that our proposed method SCORPION shows superior performance under all levels of network loss rates. An increase in the network loss rates leads to a rapid and sharp decrease in the perception performance for baseline methods, as perceptual information gets lost in the shared feature maps. For instance, with 40% network loss, SOTA methods like V2X-ViT and CoAlign see reductions in AP@0.5 of more than 20% for V2XSet and OPV2V, and over 5% in DAIR-V2X, compared to ideal communication conditions. Our method consistently outperforms previous SOTA models in all levels of network losses, clearly demonstrating the robustness of SCORPION to network loss. Note that for higher network loss, our method maintains higher improvements in accuracy thanks to the L-BEV-R module. When the network loss rate gets as high as 80%, our method outperforms the second-best approach by 5.1%, 7.5% and 3.1% on V2XSet, OPV2V, and DAIR-V2X dataset.

Results under Various Localization Errors. To demonstrate the robustness under localization error, we vary the localization noise from 0 to 1.6 m to the agents' poses in the datasets (2x higher than pervious work [13], [2], [5]). As depicted in Fig. 7, our robust perception design surpasses state-of-the-art methods that are pose-error aware, enhancing average perception performance by 6.8%, 1.3%, and 1.0% across the datasets, respectively. Comparing to SCOPE [13], a SOTA baseline that also tackles spatial and temporal misalignments, SCORPION shows 5.5% to 16.9% performance under higher noise conditions because of the DSCA models global feature interaction more precisely. Unlike CoAlign [5], which relies on prior training of an uncertainty estimator for bounding boxes, our end-to-end model with the DSCA approach offers a simple yet effective solution for addressing localization errors, thanks to the deformable attention interaction with offset locations in the misaligned feature maps.

Results under Various Synchronization Errors. The third row of Fig. 7 illustrates the perception performance under different levels of synchronization errors. Compared to methods like V2X-ViT which have designs to handle asynchronous time delay, our approach can still improve the perception accuracy by 1.1% to 11.6% across datasets. The performance boost is due to two key factors: (1) TA module aligns historical feature information to the current ego timestamp, mitigating the impact of asynchronous feature interaction from other agents and allowing feature interaction to focus on perceptually important areas from ego space; (2) DSCA module facilitates perceptually critical information interaction across agents under both spatial-temporal domains, making the perception results more robust.

D. Ablation Study

Contribution of Major Components. To evaluate the effectiveness of our proposed component, we progressively remove the three key elements in our framework. Table IV details the contribution of each component. The consistent drop in detection accuracy demonstrates the effectiveness

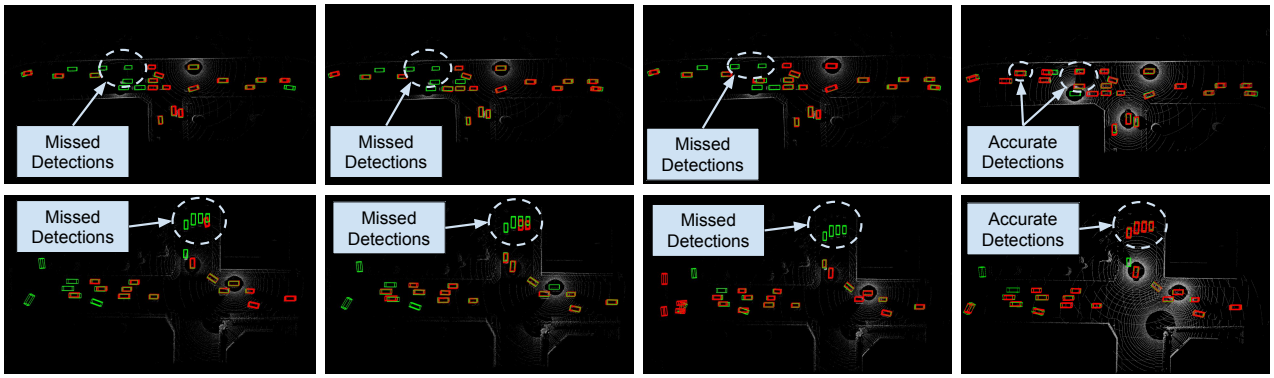

Fig. 8. Qualitative comparison results on the V2XSet dataset under **localization error, synchronization error, and network loss altogether**. The green and red boxes are ground truth and prediction outputs, respectively.

TABLE IV

FEATURE COMPARISON AND PERFORMANCE METRICS ON V2XSET. FOR BASELINE W/O DSCA, WE USE ATTENTION FUSION FROM [2].

DSCA	TA	L-BEV-R	Lossy Network	AP0.5 / AP0.7
✓	✓		N	**88.32**	**77.78**
✓			N	86.59	75.48
			N	86.09	73.52
✓	✓	✓	Y	**86.01**	**70.83**
✓	✓		Y	83.82	69.83
✓			Y	81.79	69.51
			Y	79.80	64.94

TABLE V

ABLATION STUDY RESULTS OF SCORPION (AP@0.5/0.7).

Dataset	**DAIR-V2X**	**V2XSet**	**OPV2V**
Effect of offset point numbers in DSCA			
2	89.25/78.82	91.25/77.30	71.94/53.42
4 (Default)	88.32/77.78	93.10/85.10	74.65/56.76
8	87.68/75.29	92.33/81.12	70.91/50.00
16	88.30/75.61	91.78/82.26	70.33/51.29
Effect of historical ego frame counts in TA			
1 (Default)	88.32/77.78	93.10/85.10	74.65/56.76
2	87.98/77.98	92.76/84.54	73.24/54.72
3	88.30/77.43	92.33/84.27	71.96/54.36

of each component: L-BEV-R, DSCA, and TA. By adding DSCA, AP@0.7 is improved by 2% compared to using naive attention mechanism [2]. By further incorporating past ego-agent features and performing temporal alignment, we further improve AP@0.7 by another 2.5%, indicating the effectiveness of both the DSCA and TA modules. Furthermore, incorporating the L-BEV-R module enhances SCORPION's robustness under lossy communication conditions [32], improving AP@0.5 by 2.2% compared to baselines that do not include feature reconstruction capabilities.

Impact of Offset Point Count in DSCA. We adjust the deformable offset points from 2 to 16 in DSCA (Tabel V). Noticeably, DSCA with 4 points achieves better detection performance on all three datasets. This result shows that SCORPION effectively extracts essential spatial semantics by setting a reasonable offset point number.

Effect of Historical Ego Frame Count. As shown in Table V, we evaluate the impact of incorporating different numbers of historical information from the ego agent. The finding is that using 1 historical frame is already beneficial in achieving good performance gains over all datasets.

Visualization of Detection Results. To illustrate the perception capability of different models, we provide a detection visualization[1] comparison (Fig. 8) under 1.3 m of localization error, 100 ms of asynchrony latency, and 35% of network loss based on the existing measurement work on GPS [45] and V2X network loss [32]. Baseline methods like SCOPE, V2X-ViT, and CoAlign fail to accurately detect moving objects because of the presence of three types of errors. Our proposed SCORPION handles the network packet loss through feature reconstruction and exploits the spatial and temporal misalignments through the DSCA and TA module, thus improving the detection accuracy.

V. CONCLUSION

This paper presents SCORPION, which expands AI's role in co-designing the hardware-network-AI software stack by addressing persistent issues like lossy communication and sensor noise in networking and hardware. SCORPION demonstrates improved accuracy and robustness across extensive experiments, including normal and challenging scenarios that expose typical system issues. Future work will extend SCORPION to more complex perception scenarios and incorporate additional sensor modalities, such as monocular and stereo cameras.

REFERENCES

[1] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.

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

[3] Z. Zhang and J. F. Fisac, "Safe occlusion-aware autonomous driving via game-theoretic active perception," *arXiv preprint arXiv:2105.08169*, 2021.

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

[5] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4812–4818.

[1]We also provide a video demonstration of SCORPION at https://www.youtube.com/watch?v=6zCAlhH7pGw.

[6] Q. Zhang, X. Zhang, R. Zhu, F. Bai, M. Naserian, and Z. M. Mao, "Robust real-time multi-vehicle collaboration on asynchronous sensors," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023, pp. 1–15.

[7] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.

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

[9] R. Zhu, X. Zhu, A. Zhang, X. Zhang, J. Sun, F. Qian, H. Qiu, Z. M. Mao, and M. Lee, "Boosting collaborative vehicular perception on the edge with vehicle-to-vehicle communication," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 2024, pp. 141–154.

[10] Q. Chen, X. Ma, S. 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 *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.

[11] M. Naserian, F. Bai, X. Zhang, R. Zhu, Q. Zhang, X. Zhu, and Z. M. Mao, "Cooperative v2x sensor sharing," Dec. 21 2023, uS Patent App. 17/844,978.

[12] R. Zhu, M. Cho, S. Zeng, F. Bai, X. Gao, and Z. M. Mao, "Scalable crowd-sourced global hd map construction via collaborative map perception and sparse graph fusion," in *The 4th Workshop on Transformers for Vision (T4V) at CVPR 2025*, 2025.

[13] K. e. a. Yang, "Spatio-temporal domain awareness for multi-agent collaborative perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 383–23 392.

[14] S. Hong, Y. Liu, Z. Li, S. Li, and Y. He, "Multi-agent collaborative perception via motion-aware robust communication network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 301–15 310.

[15] "OXTS Inertial Navigation Systems." https://www.oxts.com/, 2024.

[16] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun, "Learning to communicate and correct pose errors," in *Conference on Robot Learning*. PMLR, 2021, pp. 1195–1210.

[17] Y. Park, T. Kim, and D. Hong, "Resource size control for reliability improvement in cellular-based v2v communication," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 379–392, 2018.

[18] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for lte v2v communication systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3850–3860, 2018.

[19] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[20] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[21] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 4, pp. 2650–2660, 2023.

[22] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[23] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[24] X. Zhang, A. Zhang, J. Sun, X. Zhu, Y. E. Guo, F. Qian, and Z. M. Mao, "Emp: edge-assisted multi-vehicle perception," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 545–558. [Online]. Available: https://doi.org/10.1145/3447993.3483242

[25] S. Shi, J. Cui, Z. Jiang, Z. Yan, G. Xing, J. Niu, and Z. Ouyang, "Vips: Real-time perception fusion for infrastructure-assisted autonomous driving," in *Proceedings of the 28th annual international conference on mobile computing and networking*, 2022, pp. 133–146.

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

[27] M. Cho, Y. Cao, J. Sun, Q. Zhang, M. Pavone, J. J. Park, H. Yang, and Z. M. Mao, "Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion," *arXiv preprint arXiv:2410.12592*, 2024.

[28] J. B. Kenney, "Dedicated short-range communications (dsrc) standards in the united states," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1162–1182, 2011.

[29] V. Vukadinovic, K. Bakowski, P. Marsch, I. D. Garcia, H. Xu, M. Sybis, P. Sroka, K. Wesolowski, D. Lister, and I. Thibault, "3gpp c-v2x and ieee 802.11 p for vehicle-to-vehicle communications in highway platooning scenarios," *Ad Hoc Networks*, vol. 74, pp. 17–29, 2018.

[30] J. Postel, "User datagram protocol," Tech. Rep., 1980.

[31] F. Bai and H. Krishnan, "Reliability analysis of dsrc wireless communication for vehicle safety applications," in *2006 IEEE intelligent transportation systems conference*. IEEE, 2006, pp. 355–362.

[32] F. Bai, D. D. Stancil, and H. Krishnan, "Toward understanding characteristics of dedicated short range communications (dsrc) from a perspective of vehicular network engineers," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, 2010, pp. 329–340.

[33] "C-V2X Performance Assessment Project," https://tinyurl.com/2s3f2h9u, 2019.

[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[36] S. Kumar, L. Shi, N. Ahmed, S. Gil, D. Katabi, and D. Rus, "Carspeak: a content-centric network for autonomous driving," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 259–270, 2012.

[37] "Introduction to the vehicle-to-everything communications service V2X feature in 3GPP release 14." https://tinyurl.com/5n82zfzv, 2024.

[38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[39] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[40] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[41] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[43] "Open COOperative Detection framework for autonomous driving." https://github.com/DerrickXuNu/OpenCOOD, 2024.

[44] "NVIDIA A100 Tensor Core GPU." https://www.nvidia.com/en-us/data-center/a100/, 2024.

[45] M. Specht, "Consistency analysis of global positioning system position errors with typical statistical distributions," *The Journal of Navigation*, vol. 74, no. 6, pp. 1201–1218, 2021.

[46] A. Rauch, F. Klanner, and K. Dietmayer, "Analysis of v2x communication parameters for the development of a fusion architecture for cooperative perception systems," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2011, pp. 685–690.