2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Robust Collaborative Perception against Temporal Information Disturbance

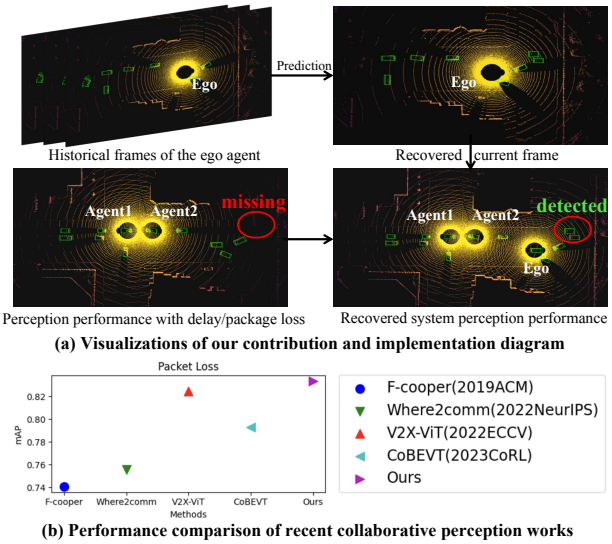
Xunjie He, Yiming Li, Te Cui, Meiling Wang, Tong Liu, Yufeng Yue*

***Abstract*— Collaborative perception facilitates a more comprehensive representation of the environment by leveraging complementary information shared among various agents and sensors. However, practical applications often encounter information disturbance which includes perception packet loss and time delays, and a comprehensive framework that can simultaneously address such issues is absent. In addition, the feature extraction process prior to fusion is not sufficient, as it lacks exploration of the local semantics and context dependencies of individual features. To enhance both accuracy and robustness, this paper introduces a novel framework named Robust Collaborative Perception against Temporal Information Disturbance, which predicts perception information when disturbance occurs. Specifically, the Historical Frame Prediction (HFP) module is introduced to make compensation for information loss with temporal association excavation of historical features. Based on the predicted features generated by the HFP module, the Pyramid Attention Integration (PAI) module is introduced to augment local semantics and incorporate global long-range dependencies through multi-scale window attention. Compared with existing methods on the publicly available dataset OPV2V, our approach exhibits superior performance and expanded robustness in the 3D object detection task. The code will be publicly available at `https://github.com/hexunjie/Ro-temd`.**

I. Introduction

Single-agent perception methods, including semantic segmentation and 3D object detection [1]–[3], have been extensively investigated. However, their effectiveness is inherently limited by single-agent constraints, such as a limited field of view, occasional obstructions, and sensor malfunctions. In response to these limitations, researchers are increasingly directing their attention towards the development of collaborative perception techniques [4]–[7] and corresponding datasets like [8]. These approaches offer advantages such as multi-perspective observation and robustness against failures in individual sensors.

Collaborative perception integrates perceptual information from diverse perspectives of multiple agents, expanding the perception range of individual agents and demonstrating significant potential for improving perception capabilities. However, multi-agent perception applications face certain degree of disturbance in practical scenarios [9], including perception packet loss and time delays. These problems essentially result in incomplete information in collaborative perception tasks. Therefore, it is necessary to compensate for the disturbed information by capturing the dynamic changes of spatial-temporal information and grasping causal relationships with the full utilization of historical features. In addition, the unified multi-agent feature shows a dense representation with rich local semantics. Directly fusing these extracted features for detection [10], [11] may lead to a loss of fine-grained texture information and long-range dependencies. The detailed excavation of local semantics and comprehensive understanding of global dependency information are worthy of research. Taking inspiration from these insights, the primary goal of this paper is to develop a novel framework that excavates temporal association features to promote robustness in practical disturbance, and enables fine-grained local semantics and global dependencies exploration of individual features.



(a) Visualizations of our contribution and implementation diagram

(b) Performance comparison of recent collaborative perception works

Fig. 1. Comparative visualization and quantification results of the proposed methods in collaborative perception scenarios with information disturbance.

The first challenge is how to make compensation for information loss to promote the robustness against for disturbance. In practical applications, perception performance can be adversely affected by information disturbance such as occasional time delays or packet loss, which is caused by communication congestion or hardware failures. However, many existing methods [5], [7], [10] assume favorable communication conditions and focus primarily on the intermediate fusion strategies. The temporal information, which represents the object appearance information and historical trending, lacks effective extraction. To address these limitations, we propose a historical frame prediction (HFP) module. It predicts and compensates for the lost features by collecting the dynamic variations and effectively capturing

*Corresponding author: Yufeng Yue (yueyufeng@bit.edu.cn).

This work was supported by the National Natural Science Foundation of China under Grant No. NSFC 62233002, 92370203.

Xunjie He, Yiming Li, Te Cui, Meiling Wang, Tong Liu, Yufeng Yue are with School of Automation, Beijing Institute of Technology, Beijing, 100081, China.

the temporal dependencies and relationships.

The second challenge revolves effectively enhancing the local semantics and global long-range dependencies of each individual agent simultaneously. The existing approaches [4], [10], [11] which directly fusing multi-perspective features without considering individual agent's feature excavation, lead to a limited comprehension of the semantic information, particularly for the fine-grained voxel features. What's more, the predicted features which preserve rich context information, lack comprehensive traversal and efficient utilization. To address this, we introduce the pyramid attention integration (PAI) module, which facilitates the acquisition of local details and global dependencies across multiple scales.

In this paper, the robust collaborative perception against temporal information disturbance network is proposed for collaborative perception. To assess perception performance in 3D recognition scenarios, the 3D object detection task has been selected as the validation benchmark. Qualitative and quantitative results demonstrate the algorithm's superior detection accuracy, as illustrated in Fig. 1. We also made the code available to the collaborative perception community. The main contributions of this paper are listed as follows:

1) The proposed network is an innovative framework, which effectively addresses challenges related to information disturbances such as perception packet loss and time delays in practical applications.
2) The HFP module is introduced to predict the current disturbed feature by excavating temporal association within historically referenced frames.
3) The PAI module is proposed to capture local semantics from multiple scales and enhance the global context correlations by shifted window design.

The remainder of this paper proceeds as follows. Section II reviews the related work. In Section III, the proposed algorithm is presented. Section IV shows the experimental analysis. Section V is the conclusion of the work.

II. Related Work

A. Collaborative Perception

Collaborative perception refers to the joint prediction achieved by multiple agents through the information sharing of diverse perspectives. This approach mitigates issues related to potential occlusion or visual distance limitations, encountered in single-agent perception scenarios. Recent efforts have been dedicated to achieving a balance between effectiveness and efficiency. In term of improving efficiency, F-cooper [4] was the pioneer in proposing a feature map fusion-based 3D object detection framework tailored for connected autonomous vehicles. Who2com [5] introduced an innovative multi-stage handshake communication mechanism, optimizing bandwidth usage by learning the most relevant communication partners. When2com [6] established communication groups and determined the most opportune times for communication. Subsequently, Glaser et al. [12] designed a smooth self-coding method based on it, incorporating an adjustable compression module to fine-tune bandwidth usage. Where2comm [7] leveraged a novel spatial confidence map at each agent to realize effective compression. V2VNet [11] devised an information transfer mechanism based on a spatial graph neural network, enabling collaborative perception and prediction in autonomous driving. In contrast to prior methods which adopts a single supervision framework, DiscoNet [10] employed a "student-teacher" knowledge distillation network to heighten direct supervision of intermediate features, facilitating efficient reasoning in multi-agent collaboration. CoBEVT [13] introduced a fused axial attention module to collaboratively generate bird's eye view (BEV) predictions using a Transformer-based approach. These multi-agent distributed collaboration works have made great progress in bandwidth occupancy and cognitive accuracy respectively. However, most of existing networks consider more about the intermediate fusion strategy, which neglect the local semantic excavation and global context dependencies of each individual agent.

B. Time Series Prediction

Time series prediction techniques [14] utilize historical information to anticipate future trends, leveraging inherent correlations across multiple dimensions. Early on, Recurrent Neural Network (RNN) methods [15], including LSTM [16] and GRU [17], were introduced for forecasting future signals. Building upon this, Wang et al. [18] proposed a predictive neural network that extracts temporal dynamic information. Su et al. [19] presented a convolutional tensor-train LSTM for spatio-temporal learning. Considering potential interruptions or packet loss in collaborative perception, time series prediction strategies play an important role in enhancing perception robustness. SyncNet [9] was introduced to mitigate the inevitable impact of time delays, adapting asynchronous collaborative features to the same timestamp for robustness improvement. St-p3 [20] designed a dual-pathway model for time domain prediction, ensuring stronger feature representation and improved semantic inference performance. BEVerse [21] generated BEV representations from multi-camera videos and jointly inferred across multiple tasks. While existing networks have made significant contributions to various tasks in robot perception, there remains a need for a framework that effectively handles common exceptional disturbances. The internal associations and interactions between historical features have not been fully explored, directly impacting the accuracy of current feature prediction and robustness in disturbed scenarios.

III. Methodology

In this section, the robust collaborative perception against temporal information disturbance network is presented and the introduction is divided into three subsections: problem formulation and architecture design, historical prediction module and pyramid attention integration module.

A. Problem Formulation and Architecture Design

We consider a collaborative perception system with N agents, where one of them is designated as the ego agent. The

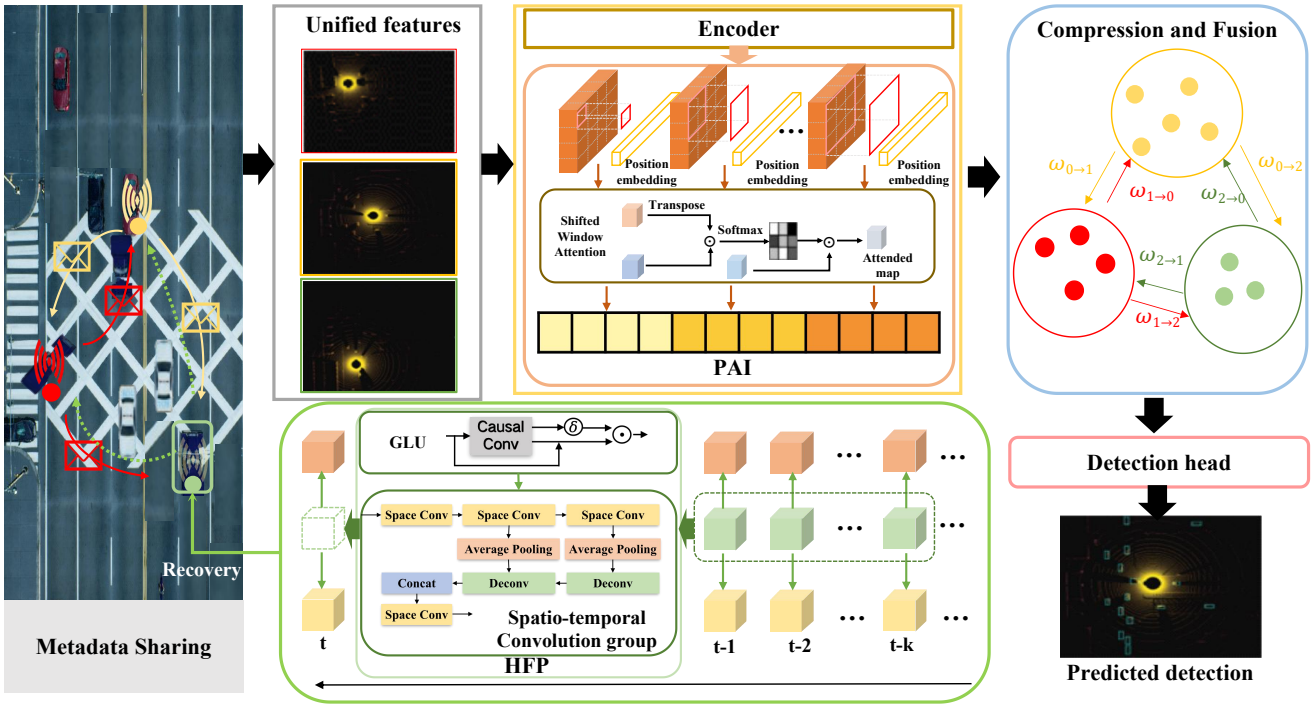


Fig. 2. The overall architecture of our method. It is divided into four parts: metadata sharing, feature encoder, compression and fusion, and detection head. HFP and PAI refer to our proposed Historical Frame Prediction and Pyramid Attention Integration module, respectively.

metadata input for each agent $i \in [1, \ldots, N]$ encompasses point cloud, pose, timestamps, and multiple frame details. Specifically, spatial features comprise point cloud and pose, while temporal features include timestamps and multiple continuous frames. The point cloud at frame t, denoted as $\mathbf{X}_i^t$, serves as the input for our perception system. Fig. 2 illustrates the overall architecture of the network, consisting of four components: metadata sharing, feature encoder, compression and fusion, and the detection head.

1) Metadata Sharing: During the metadata sharing phase, each agent exchanges pertinent information, including point clouds and poses, with other agents. Upon receiving the pose of the ego agent, all connected agents within the group use the coordinate transformation matrix to project their individual point clouds onto the ego agent's coordinate system. This procedure streamlines the extraction of unified multi-perspective features, facilitating subsequent analysis and processing. What's more, when the packet loss or delay problem occurs in practical applications, our historical frame prediction (HFP) module utilizes historical frames $[\mathbf{X}_i^{t-\tau}, \ldots, \mathbf{X}_i^{t-1}]$ to predict current frame $\mathbf{X}_i^t$.

2) Feature Encoder: The encoder based on the PointPillar baseline [22] is utilized to extract feature maps from the inputs:

$$\mathbf{F}_i^t = Enc(\mathbf{X}_i^t), \tag{1}$$

where $Enc(\cdot)$ is the encoder function. Specifically, the raw point clouds are converted into a stacked pillar tensor, which is then scattered to form a pseudo image. This pseudo image is then sent back to the original baseline. The encoder extracts feature maps at frame t denoted as $\mathbf{F}_i^t$ and transforms the feature representations to BEV styles in the same global coordinate system.

Recognizing the limitations of directly fusing features, which can result in a loss of semantic information, we propose a multi-local learning approach on the individually extracted features. The proposed pyramid attention integration (PAI) module enhances the feature mining with the output feature $\mathbf{F}_{oi}^t, i \in [1, \ldots, N]$ by capturing global and local information simultaneously.

3) Compression and Fusion: To reduce the transmission bandwidth between agents, we utilize multiple continuous 1×1 convolution layers to compress the BEV features along channel dimensions. Then, we fuse features from other agents to generate more accurate features for the ego agent:

$$\mathbf{F}_{fuse}^t = Fus(\mathbf{F}_{o1}^t, \ldots, \mathbf{F}_{oN}^t). \tag{2}$$

where $Fus(\cdot)$ is the intermediate fusion strategy. Conventionally, the ego agent acts as the central node of a graph, with other agents forming a network around it. These agents are assigned weights based on the comparison learning between features of different agents. Subsequently, a weighted summation operation is performed to aggregate the information from the surrounding agents.

4) Detection Head: The detection decoder decodes features into class and regression outputs:

$$\mathbf{Y}_r^t = Dec(\mathbf{F}_{fuse}^t), \tag{3}$$

where $Dec(\cdot)$ is the decoder function. Each location of $\mathbf{Y}_r^t$ represents a rotated box with class $(x, y, h, w, l, h, \alpha)$, in which (x, y, h) is the position, (w, l, h) is the size and

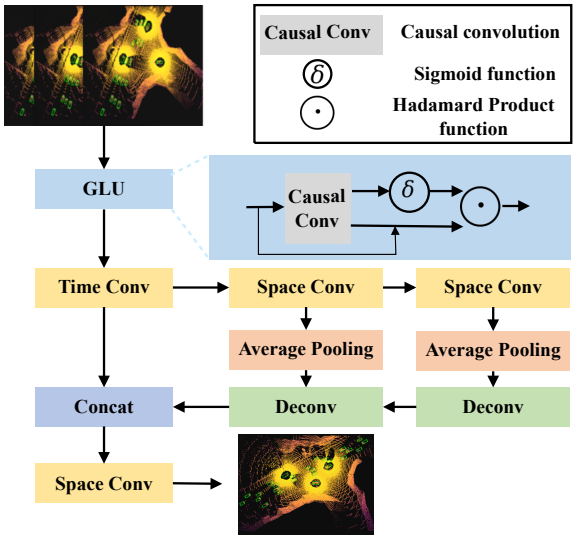


Fig. 3. The historical frame prediction module with a GLU module and a spatio-temporal convolution group.

α denotes the angle, respectively. The class output is the confidence score of an anchor box for framing a category.

B. Historical Frame Prediction (HFP) Module

In real scenarios, there may be sudden information loss or delay in the process of collaborative perception for an agent. The HFP module can generate the prediction of the current frame through a series of multidimensional convolution mixing operations, taking the stacked results of the previous few historical frames as input.

As shown in Fig. 3, the input of the HFP module $\mathbf{F}_{his}^{t} \in \mathbf{R}^{k\times C\times H\times W}$ is formed by stacking historical frames, where k is the number of historical frames required for the prediction at the current timestamp t. The module firstly uses a gated linear unit (GLU) to aggregate historical frames from time dimention, followed by a 1-D convolutional layer to generate the temporal feature $\mathbf{T}^{t}$,

$$\mathbf{T}^{t} = C_t(GLU(\mathbf{F}_{his}^{t})), \tag{4}$$

$$GLU(\mathbf{X}) = (\mathbf{X}+Caus(\mathbf{X}*W+b))\odot(\delta(Caus(\mathbf{X}*V+c))), \tag{5}$$

where $\mathbf{C}_t(\cdot)$ is the 1-D convolutional layer in time dimension. $Caus(\cdot)$ means the causal convolution where an output at time t is convolved only with elements from time t and earlier in the previous layer. W, V are the seperation weight matrices and b, c are the bias parameters. In addition, δ is a Sigmoid function and $\odot$ is a Hadamard Product function.

It is followed by a two-branch spatio-temporal convolution group. Both of the branches consist of a 2-D convolutional, a global average pooling and a deconvolution operation. Specifically, the first branch utilizes the temporal feature $\mathbf{T}^{t}$ as input, and then the output after its 2-D convolution operation serves as the input of the second branch. Sequentially, they generate feature $\mathbf{S}_1^{t} \in \mathbf{R}^{1\times 2C\times H\times W}$ and $\mathbf{S}_2^{t} \in \mathbf{R}^{1\times 2C\times H\times W}$, respectively. The connected results

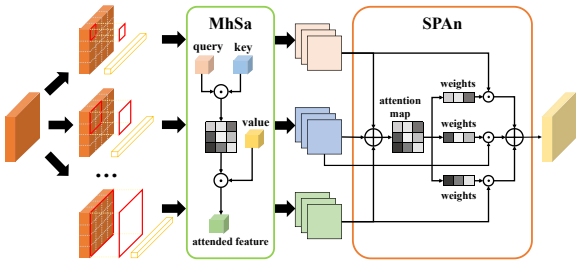


Fig. 4. The pyramid attention integration module with multi-head self-attention and space attention.

$\mathbf{F}_{concat}^{t} \in \mathbf{R}^{1\times 5C\times H\times W}$ are obtained then by a concatenation operation.

$$\mathbf{S}_1^{t} = DeC(AP_t(C_s(\mathbf{T}^{t}, \gamma))), \tag{6}$$

$$\mathbf{S}_2^{t} = DeC(DeC(AP_t(C_s(C_s(\mathbf{T}^{t}, \gamma))))), \tag{7}$$

$$\mathbf{F}_{concat}^{t} = \mathbf{T}^{t} \oplus \mathbf{S}_1^{t} \oplus \mathbf{S}_2^{t}, \tag{8}$$

where $AP_t(\cdot)$ is a global average pooling operation in time dimension, $C_s(\cdot)$ is a 2-D convolutional layer in spatial dimension, $DeC(\cdot)$ is the upsampling operation, $\oplus$ is a concatenate operation and γ is the alignment parameter.

Finally, through the 2-D convolutional layer, we obtain the prediction of a current lost frame $\mathbf{F}_{predict}^{t} \in \mathbf{R}^{C\times H\times W}$.

$$\mathbf{F}_{predict}^{t} = C_s(\mathbf{F}_{concat}^{t}) \tag{9}$$

Compared to simple time dimension convolution and pooling techniques, our proposed HFP module has been enhanced and optimized with a GLU module and a spatio-temporal convolution group. Specifically, spatial dimension convolution and deconvolution operations have been incorporated, where deeper features have been explored to capture richer and more informative representations. What's more, the causal convolution in the GLU module makes full use of continuous information in time dimension.

C. Pyramid Attention Integration (PAI) Module

Drawing inspiration from the Swin Transformer [23], we incorporate a pyramid attention integration module in our approach to enhance the internal interaction of features in the spatial dimension prior to intermediate fusion. As depicted in Fig. 4, the pyramid attention integration module divides the features into different branches based on window sizes of varying scales. Each branch then undergoes a window self-attention operation. Subsequently, the feature outputs from all branches are adaptively fused using a spatial attention module.

This pyramid attention integration module effectively captures and integrates information at multiple spatial scales, enabling improved feature representation and interaction. Specifically, large-scale window attention effectively encompasses a broader context and mitigates localization errors, while small-scale window attention enhances intrinsic correlations and magnifies local semantic information.

Specifically, with the extracted middle feature $\mathbf{F}_{i}^{t} \in \mathbf{R}^{B\times L\times H\times W\times C}$, each branch converts it into window features of different shapes based on different scales

TABLE I

COMPARATIVE RESULTS IN DIFFERENT PACKET LOSS RATES OBTAINED ON THE PUBLIC OPV2V DATASET [8].

Networks	0%	20%	40%	60%	80%	AVG
F-Cooper [4]	0.7772	0.7629	0.7493	0.7324	0.7172	0.7404
Where2comm [7]	0.7291	0.7472	0.7607	0.7466	0.7668	0.7557
V2X-ViT [24]	0.8594	0.8435	0.8291	0.8204	0.8058	0.8248
CoBEVT [13]	0.8223	0.8097	0.7988	0.7881	0.7767	0.7930
Ours	**0.8735**	**0.8578**	**0.8430**	**0.8258**	**0.8100**	**0.8337**

TABLE II

RESULTS OF DIFFERENT LEVELS OF DELAY (MILLISECOND).

Networks	0	100	200	300	400	AVG
F-Cooper [4]	0.7772	0.7721	0.7360	0.6511	0.6069	0.7087
Where2comm [7]	0.7291	0.7255	0.6913	0.6134	0.5695	0.6658
V2X-ViT [24]	0.8594	0.8294	0.7957	0.7681	0.7406	0.7986
CoBEVT [13]	0.8223	0.8158	0.7755	0.6956	0.6487	0.7516
Ours	**0.8735**	**0.8707**	**0.8510**	**0.7981**	**0.7501**	**0.8287**

$[s_1, s_2, \ldots, s_m]$, where m is the number of branches. The window size and the window feature of the $j-th$ branch are $s_j = j \times s_1$ and $\mathbf{W}^p \in \mathbf{R}^{B\times L\times n_p\times(\frac{H}{s_j}\cdot\frac{W}{s_j})\times(s_j\cdot s_j)\times C}$, where $p = [1, 2, \ldots, \frac{H}{s_j}\cdot\frac{W}{s_j}]$ and n_p denotes the number of heads. Then each branch performs multi-head self-attention on the window features and then converts them back to $\mathbf{F}^j_{mhsa} \in \mathbf{R}^{B\times L\times H\times W\times C}$. At the same time, a position embedding on the weight map is added in multi-head self-attention to further improve positioning accuracy as (11).

$$\mathbf{F}^j_{mhsa} = \Phi(MhSa(\mathbf{W}^p)), \tag{10}$$

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d} + B)V, \tag{11}$$

where $\Phi(\cdot)$ means an rearrange operation and $MhSa(\cdot)$ is the internal window attention function with the transformer principle as (11). B, Q, K, V and d refer to the relative position embedding, query, key and value matrices and the query dimension, respectively.

Finally, the PAI module generates attention weights for each feature from the channel dimension and performs a weighted summation to adaptively fuse the features from each branch:

$$\mathbf{F}^t_{oi} = S_{PAn}(\mathbf{F}^1_{mhsa}, \ldots, \mathbf{F}^m_{mhsa}), \tag{12}$$

$$S_{PAn}(\mathbf{X}^1, \ldots, \mathbf{X}^m) = C_{on}(\Sigma^q_{u=1}\omega_u\mathbf{X}^1_u, \ldots, \Sigma^q_{u=1}\omega_u\mathbf{X}^m_u), \tag{13}$$

$$\omega_u \in \Omega = AP(\mathbf{X}), u \in [1, q], \tag{14}$$

where $C_{on}(\cdot)$ and $AP(\cdot)$ refer to the concatenation and global average pooling operation. Ω and q are the weight matrix and channel number of each feature, respectively.

IV. EXPERIMENTAL RESULTS

In this section, the proposed algorithm is compared with the state-of-the-art algorithms. They are evaluated with extensive experiments on the open public collaborative perception dataset OPV2V [8].

A. Dataset

OPV2V is a large-scale Vehicle-to-Vehicle perception dataset co-simulated by CARLA and OpenCDA, which has on average approximately 3 connected vehicles in each frame. Specifically, approximately 55% frames are contained in the training set, whereas the validation and test sets include 25% and 20% frames, respectively. The dataset includes 11464 frames of 3D point clouds and 232913 annotated 3D boxes.

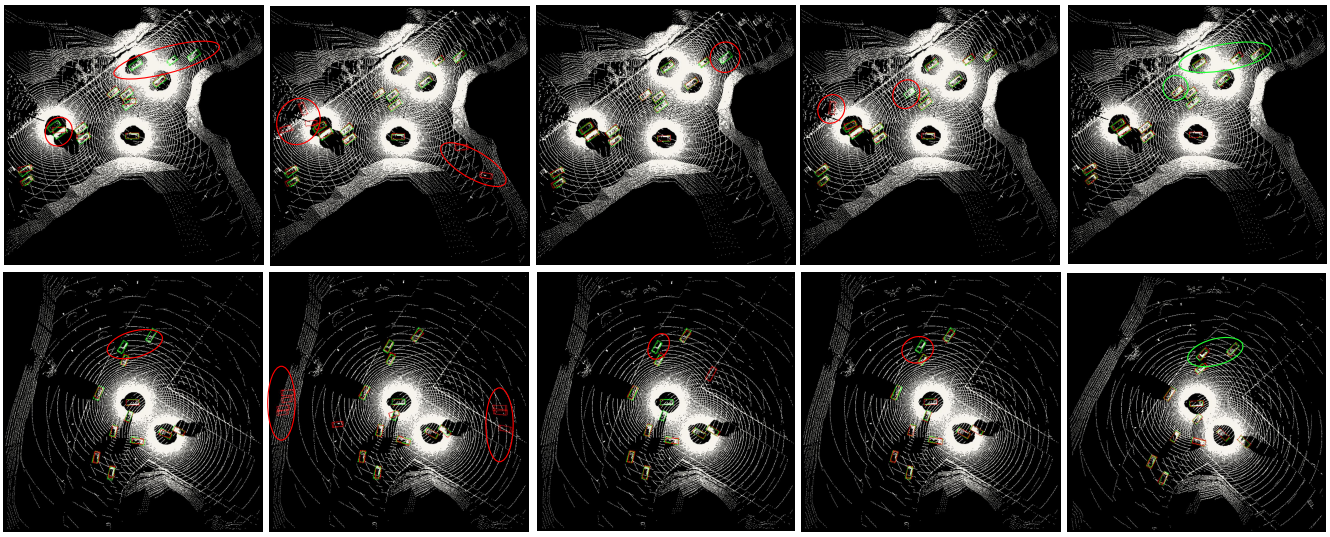
B. Implementation Details

Before training, the width/length of each voxel is set as $0.4m$, and the height is $4m$. We employ the conventional BEV detection evaluation metric: mean Average Precision (mAP). During training, the learning rate is initialized as 0.00002 and the weight decay is set as 0.0001. We set the batch size as 2. The optimization solver is Adaptive Moment Estimation (Adam). Then, we target the detection of the car category and report the results on the test set. In addition, our network is trained with a single NVIDIA RTX 3090 graphics card.

C. Comparative Results

1) Baselines and existing methods: Considering about existing perception methods, early collaboration with raw data and late collaboration which applies a global non-maximum suppression (NMS) both work in multiple agents perception. In terms of intermediate fusion, we consider F-Cooper [4], V2X-ViT [24], CoBEVT [13] and Where2comm [7]. All the methods above are proposed based on the same detection architecture with PointPillar as the backbone, and collaborate with each other at the same intermediate feature layer. The visual comparison of various methods under the condition of lossless testing is illustrated in Fig. 5. Compared to other networks, our method achieves the most comprehensive and precise detection results, demonstrating superior performance.

2) Comparison in different packet loss rates: Considering the occurrence of random information loss, we randomly apply noise to a random agent in a certain frame. Specifically, the loss rate denotes the percentage of information loss of the frame. Our method achieves remarkable performance, as demonstrated by the superior results and strong performance across different rates in Tab. I, when compared to existing open methods. Notably, our proposed network exhibits minimal performance degradation even with the same degree of information loss. This resilience is attributed to the real-time compensation provided by the HFP module, enabling efficient prediction of lost current frame features.



(a) F-cooper (b) Where2comm (c) V2X-ViT (d) CoBEVT (e) Ours

Fig. 5. Qualitative comparison between our method and existing networks on OPV2V dataset [8]. Green and red 3D bounding boxes represent the ground truth and prediction respectively. The missed or incorrect detections are highlighted in red circle while parts of the correct ones are highlighted in green.

3) Comparison in different levels of delay: In practical communication, time delays are a common occurrence that can negatively affect detection performance by introducing incorrect temporal information. To evaluate the effectiveness of our method, we conduct a comparison with existing open-source approaches, and the results are summarized in Tab. II. We consider delays $(millisecond)$ at various levels, and the detection accuracy indirectly reflects the prediction capability of our work. It is evident that our method consistently outperforms other methods across different delay levels, demonstrating its significant impact in mitigating the effects of abnormal delay conditions.

4) Comparison in different location errors: Considering the potential presence of localization errors in real-life scenarios, we design experiments at different error levels. The noise $(meter)$ is added on x, y axis according to the groundtruth position, which is simulated by Gaussian distribution. By comparing our method with other networks, we validate the robustness of our approach as illustrated in Tab. III. The experimental results showcase our best performance, demonstrating excellent generalization capbility and fault tolerance in the presence of localization noise.

TABLE III

RESULTS OF DIFFERENT LOCALIZATION ERRORS (METER).

Networks	0	0.2	0.5	0.8	AVG
F-Cooper [4]	0.7772	0.7767	0.7558	0.7376	0.7618
Where2comm [7]	0.7291	0.7404	0.7347	0.7008	0.7263
V2X-ViT [24]	0.8594	0.8540	0.8270	0.7856	0.8315
CoBEVT [13]	0.8223	0.8210	0.8198	0.8106	0.8184
Ours	**0.8735**	**0.8688**	**0.8556**	**0.8258**	**0.8559**

TABLE IV

AN ABLATION STUDY THAT INVOLVES THE INTEGRATION OF DIFFERENT COMPONENTS IN OUR METHOD.

Variants	mAP@0.3	mAP@0.5
w/o HFP, w/o PAI	0.7291	0.7008
w/o HFP	0.7824	0.7547
w/o PAI	0.7932	0.7498
Ours	**0.8735**	**0.8545**

D. Ablation Study

To evaluate the effectiveness of the proposed two modules, we conduct an ablation study as shown in Tab. IV. Specifically, 'Without' is abbreviated to 'w/o'. 'w/o HFP, w/o PAI' means the baseline of our work which removes the HFP module and the PAI module. Starting from the baseline, the network without HFP module is denoted as 'w/o HFP'. The 'w/o PAI' variant with the disturbed input is analogously introduced. Ours achieves the best performance with two novel modules, which both works better than the baseline. What's more, 'w/o PAI' works better than 'w/o HFP', which demonstrates that the HFP module with historical information contributes more to the end result.

V. CONCLUSION

We propose an innovative framework for collaborative perception in the presence of disturbance issues such as packet loss or time delays, specifically focus on enhancing both accuracy and robustness in 3D object detection. The framework introduces an HFP module that utilizes historical continuous features to predict the abnormal features, effectively compensating information deviation. Additionally, a specially designed PAI module is incorporated to extract local semantics and global dependencies from individual agent features and facilitate optimal utilization of the predicted features. Thorough comparisons and detailed studies show that our method performs better than existing approaches. In the future, we plan to expand the collaborative perception framework to work with different modalities of data, making it more versatile for various visual tasks.

REFERENCES

[1] X. He, M. Wang, T. Liu, L. Zhao, and Y. Yue, "Sfaf-ma: Spatial feature aggregation and fusion with modality adaptation for rgb-thermal semantic segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[2] M. Wang, L. Zhao, and Y. Yue, "Pa3dnet: 3-d vehicle detection with pseudo shape segmentation and adaptive camera-lidar fusion," *IEEE Transactions on Industrial Informatics*, 2023.

[3] L. Zhao, M. Wang, and Y. Yue, "Sem-aug: Improving camera-lidar feature fusion with semantic augmentation for 3d vehicle detection," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9358–9365, 2022.

[4] 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.

[5] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.

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

[7] 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.

[8] 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.

[9] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *European Conference on Computer Vision*. Springer, 2022, pp. 316–332.

[10] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[11] 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.

[12] N. Glaser, Y.-C. Liu, J. Tian, and Z. Kira, "Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2406–2413.

[13] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative birdâs eye view semantic segmentation with sparse transformers," in *Conference on Robot Learning*. PMLR, 2023, pp. 989–1000.

[14] Z. Han, J. Zhao, H. Leung, K. F. Ma, and W. Wang, "A review of deep learning models for time series prediction," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7833–7848, 2019.

[15] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[16] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International conference on big data (Big Data)*. IEEE, 2019, pp. 3285–3292.

[17] X. Li, X. Ma, F. Xiao, C. Xiao, F. Wang, and S. Zhang, "Time-series production forecasting method based on the integration of bidirectional gated recurrent unit (bi-gru) network and sparrow search algorithm (ssa)," *Journal of Petroleum Science and Engineering*, vol. 208, p. 109309, 2022.

[18] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9154–9162.

[19] J. Su, W. Byeon, J. Kossaifi, F. Huang, J. Kautz, and A. Anandkumar, "Convolutional tensor-train lstm for spatio-temporal learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 714–13 726, 2020.

[20] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 533–549.

[21] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[22] 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.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[24] 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.