Robust Collaborative Perception against Communication Interruption

Shunli Ren , **Zixing Lei** , **Zi Wang** , **Siheng Chen** , **Wenjun Zhang**
Shanghai Jiao Tong University
{renshunli,chezacarss,w4ngz1,sihengc,zhangwenjun}@sjtu.edu.cn

Abstract

Collaborative perception has attracted much attention because it effectively improves the perception performance beyond the limited perception ability of the individual agent. However, most previous works only consider ideal communication among agents without interruption, which could seriously affect collaboration performance. To alleviate the effect of communication interruption, we propose a novel interruption-aware robust collaborative perception (IA-RCP) framework, which leverages historical information to recover missing information due to the communication interruption. To further improve recovery performance, we design a trainable spatial attention mask to suppress background noise and a curriculum learning strategy to stabilize training. Experiments demonstrate that our method can bring significant benefits to alleviate the effect caused by communication interruption.

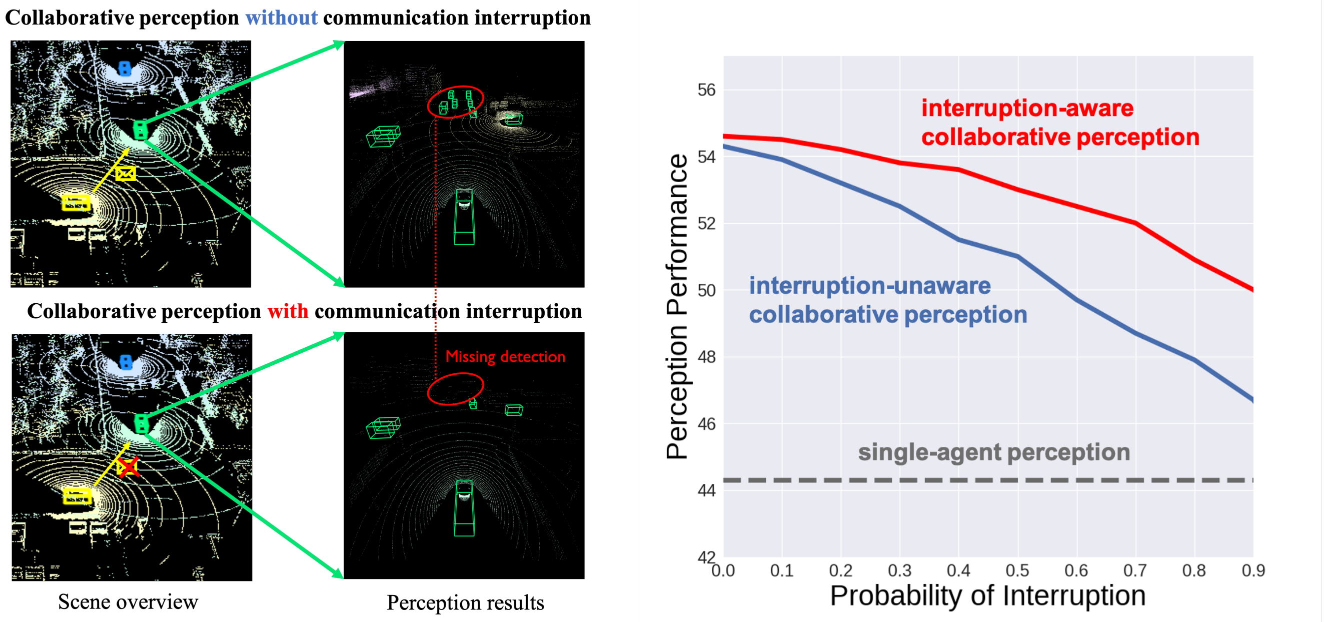
1 INTRODUCTION

Perception is essential for various robotic systems and has attracted a tremendous amount of attention from various fields [Ren *et al.*, 2015; Shi *et al.*, 2019; Chen *et al.*, 2018]. With the rapid developments of advanced sensors and algorithms, the perception ability has made great progress during the past decade. However, single-agent perception is fundamentally limited due to a constrained perception range. For example, for a single agent, long range and occlusion scenarios are almost impossible to be solved. To address these issues, collaborative perception has been proposed to enable neighboring agents to share information with each other; so that each agent can perceive the surrounding environment beyond line-of-sight and field-of-view. Related techniques are useful in a wide range of real-world applications, such as vehicle-to-everything-communication-aided autonomous driving [Wang *et al.*, 2020; Li *et al.*, 2021], multi-robot warehouse automation system [Li *et al.*, 2020; Zaccaria *et al.*, 2021] and multi-UAVs (unmanned aerial vehicles) for search and rescue [Scherer *et al.*, 2015; Alotaibi *et al.*, 2019]. Recently, some works [Chen *et al.*, 2019; Miller *et al.*, 2020; Arnold *et al.*, 2020] adopted this idea, proposed several effective collaborative perception methods, and validated the effectiveness of collaborative perception.

The current success of collaborative perception depends on not only a well-designed collaboration strategy, but also ideal communication conditions. Unfortunately, real-world communication is rarely perfect. Even though communication technology is developing explosively, some fundamental issue is still inevitable. For example, *random temporary interruption* is one of the common communication problems caused by environmental factors such as unstable communication channels and equipment failure. In this case, every communication link between two agents could be interrupted with a certain probability at each moment. This results in a dynamic, incomplete communication graph, which would severely degrade the collaboration performance and further affect the downstream tasks, such as tracking and trajectory prediction, causing a cascading failure; see Figure 1.

To fill this gap, this work considers promoting robust collaboration gain even with stochastic communication interruption. It is worth noting that our specific aim is to alleviate the effect when interruption happens from an perspective of machine learning algorithm; instead of how to avoid the interruption from a communication perspective. To achieve this goal, we propose an interruption-aware robust collaborative perception (IA-RCP) framework, which leverages historical information to recover the missing information due to the communication interruption. To achieve reliable recovery, the proposed IA-RCP framework has two key features. First, we bridge historical information to missing information through spatial-temporal correspondence. Since the interruption randomly happens, one agent may receive the relevant information from the currently disconnected agents in history. Thus, we need to infer the current missing messages from history according to the spatial-temporal relationship to alleviate the effect of the missing information. Second, we conduct missing information recovery in the intermediate feature domain instead of converting to raw data space. This allows easier and more direct feature extraction and usage.

To further improve the performance of IA-RCP, we propose a trainable spatial attention mask constraint to suppress the noise and error generated from inaccurate prediction. We also adopt a curriculum learning strategy to promote more stable training. The training process starts with low interruption probability, and the range of the interruption probabil-



(a) Interruption issue.

(b) Affect of perception performance.

Figure 1: Communication interruption issue and its affect. Plot (a) shows that the green vehicle can expand its perception range to overcome occlusion and long-range issues and detect more objects by leveraging the supportive message sent by the yellow vehicle. However, when the communication between two vehicles is stochastically interrupted, the performance of collaborative perception becomes unstable: the detected objects sometimes appear, sometimes are missing, causing noisy inputs for the downstream tasks, such as tracking and trajectory prediction. Plot (b) shows the empirical performances of collaborative perception methods as a function of interruption probability on the V2X-Sim [Li *et al.*, 2021] dataset. We see that the performance are seriously degraded due to communication interruption. Fortunately, the proposed interruption aware collaborative perception framework (red curve) effectively alleviates the degradation.

ity increase gradually. We validate our method on the public large-scale collaborative perception dataset V2X-Sim [Li *et al.*, 2021]. The results show that our IA-RCP framework can improve the performance of other existing collaboration strategies at various interruption probabilities and effectively alleviate the effect of the communication issue. The maximum improvement is up to 4.47%.

To summarize, our main contributions are as follows:

- We propose an interruption-aware robust collaborative perception (IA-RCP) framework, leveraging historical information to recover missing information due to communication interruption. To our best knowledge, this is the first work to address the interruption issue in collaborative perception;
- We propose two further designs to improve the quality of missing information recovery: spatial attention mask constraint for background suppression, and curriculum learning strategy for more stable training;
- We conduct extensive experiments on V2X-Sim dataset to show that the proposed IA-RCP brings significant benefits to alleviate the effect of communication interruption.

2 RELATED WORK

2.1 Collaborative Perception

Collaborative perception can dramatically improve the perception performance compared with single-agent perception and overcome the physical limitations of single-agent perception with limited sensor ability, such as the occlusion and long-range issue. In the image segmentation task, [Liu *et al.*, 2020b] introduced a handshake mechanism to determine which agent should communicate with; To decide the communication timing, [Liu *et al.*, 2020a] proposed an asymmetric attention mechanism which establishes communication groups also in image segmentation task; In 3D autonomous driving scenarios, [Wang *et al.*, 2020] proposed a multi-round communication mechanism for joint perception and prediction task; [Li *et al.*, 2021] introduced a collaborative algorithm using knowledge distillation technology to reduce bandwidth consumption. It also generates point-wise fusion weight for each spatial coordinate to improve performance.

Most of the previous works discussed the communication strategy under an assumed perfect communication system. However, in realistic scenarios, communication is never perfect. This work considers the effect of communication interruption, a common anomaly in the communication process, in collaborative perception and proposes a robust model for collaborative perception against communication interruption.

2.2 Vehicle-to-vehicle (V2V) Communication

Vehicle-to-vehicle (V2V) communication can be implemented by two communication solutions, either IEEE 802.11p protocol or cellular network standards [Mei *et al.*, 2018]. In IEEE 802.11p protocol, stations do not need to join a BSS (Basic Service Set) by operating in WAVE (Wireless Access in Vehicular Environment) mode, which reduces the connection setup overhead and suits vehicular safety applications well [Jiang and Delgrossi, 2008]. On the other hand, the fourth-generation cellular networks support LTE V2V standard development, supporting vehicular user equipments (VUEs) with low latency and highly reliable data transmission [Araniti *et al.*, 2013]. Compared to the 802.11p based V2V communication, it avoids channel congestion and collision induced by CSMA mechanism [Lei *et al.*, 2016]. Though communication technology keeps developing for lower latency and better reliability, some fundamental problem like interruption will inevitably exist for a long time. Our work in this paper is to alleviate the effect of interruption from an machine learning perspective instead of avoiding it from a communication perspective.

3 Problem Formulation

We consider there are N agents in a collaborative perception system. Each agent a_i is provided with the accurate pose information and has a set of neighbour agents $\mathcal{N}_i$. At each time stamp t, agent a_i observes its surrounding environment to obtain local observation $\mathbf{X}_i^{(t)}$. Then agents use an encoder f_{encode} to exact features of $\mathbf{X}_i^{(t)}$ and produce the messages to be transmitted $\mathbf{F}_i^{(t)}$, that is,

$$\mathbf{F}_i^{(t)} = f_{\text{encode}}(\mathbf{X}_i^{(t)}).$$

Once produced, the features will be transmitted to neighbouring agents, and each agent receives the features from its neighbours. After that, each agent a_i fuses its own observation and received features with a fusion model f_{fuse} to produce a fused feature $H_i^{(t)}$, that is

$$\mathbf{H}_i^{(t)} = f_{\text{fuse}}(\{\mathbf{F}_j^{(t)}\}_{j\in\{i\}\cup\mathcal{N}_i}) \qquad (1)$$

Finally, a decoder f_{decode} is used to decode the fused features and obtain the perception results $\widehat{\mathbf{Y}}_i^{(t)}$ of each agent as follows,

$$\widehat{\mathbf{Y}}_i^{(t)} = f_{\text{decode}}(\mathbf{H}_i^{(t)}).$$

The ground truth of the perception results of agent a_i at time stamp t is denoted by $\mathbf{Y}_i^{(t)}$.

Stochastic communication interruption. When the stochastic communication interruption is taken into consideration, each communication link between any two agents may interrupted with a certain probability so that agents will not be able to receive messages from all neighbour agents. As shown in Figure 1(b), perception performance will be significantly degraded by the communication interruption. Therefore, our specific aim is to make collaborative perception performance robust to the loss of messages caused by stochastic communication interruption.

4 METHODOLOGY

This section presents the proposed interruption-aware robust collaborative perception(IA-RCP) framework. We first introduce missing information recovery based on historical information. We then aggregate the recovered information with the ego information and the other well-received information to achieve comprehensive fusion. We further present two advanced designs to improve the IA-RCP framework. Finally, we show the training loss.

4.1 Missing Information Recovery

The communication interruption happens randomly. Though agent a_i fails to receive messages from agent a_j at time t, it may have received the relevant information from a_j or other agents in the past. Motivated by this, we propose to infer the current missing information based on historical information through a missing information recovery process. The recovery process is composed of a completion model and a prediction model. The completion model recovers missing information at each historical time stamp. The prediction model then recovers the current feature based on the completed features in the historical frames. The overall process is shown in the top row of Figure 2. The detail of each model will be introduced in the following.

Historical information. Here, we assume that each agent a_i stores the fused features of the past k key frames, denoted by $\mathbf{H}_i^{(t-\tau)}$, $\tau = 1, 2, \cdots, k$. Note that $\mathbf{H}_i^{(t-\tau)}$ is fused by the incomplete features $\{\mathbf{F}_j^{(t-\tau)}\}_{j\in\{i\}\cup\mathcal{R}_i^{(t-\tau)}}$, where $\mathcal{R}_i^{(t)} \in \mathcal{N}_i$ is the set of neighbour agents from which agent a_i can receive messages at time stamp t. The received unfused features are aggregated to obtain the fused feature $\mathbf{H}_i^{(t-\tau)}$ and the fusion weight of each unfused features $\mathbf{M}_j^{(t-\tau)}$ at each time stamp, which provide us the temporal and spatial information from history. As shown in the left top part of Figure. 2, we transform all features in history to the coordinate system of current time t based on the pose information to obtain the historical features; that is,

$$\{\mathbf{H}_i^{(t-\tau\to t)}\}_{\tau=1}^k = \xi^{(t-\tau\to t)}(\{\mathbf{H}_i^{(t-\tau)}\}_{\tau=1}^k),$$

where the subscript $(t-\tau \to t)$ represents the coordinate transformation from the coordinate system of time $t-\tau$ to that of time t and $\xi^{(t-\tau\to t)}$ is the transformation principle based on the poses at two time points.

Completion model. Since communication interruption happens stochastically at each time stamp, the historical information may not be complete as well. Intuitively, agent a_i could get information from both a_j and and a_k at time t, get information only from a_j at time $t+1$, and get information only from a_k at time $t+2$. This causes collaboration information temporally inconsistent, which makes the recovery of current missing information difficult.

To solve this issue, we propose the completion model to make the features at different times consistent with each other, illustrated in Figure 3. From the feature at first $\mathbf{H}_i^{(t-k\to t)}$, we predict its state at the next time stamp and fuse it with the feature received at next time stamp $t-k+1$ to obtain the feature after completion at $t-k+1$, $\mathbf{Z}_i^{(t-k+1\to t)}$. We repeat this process until the last history feature at $t-1$. Each step of the process is formulated as follows,

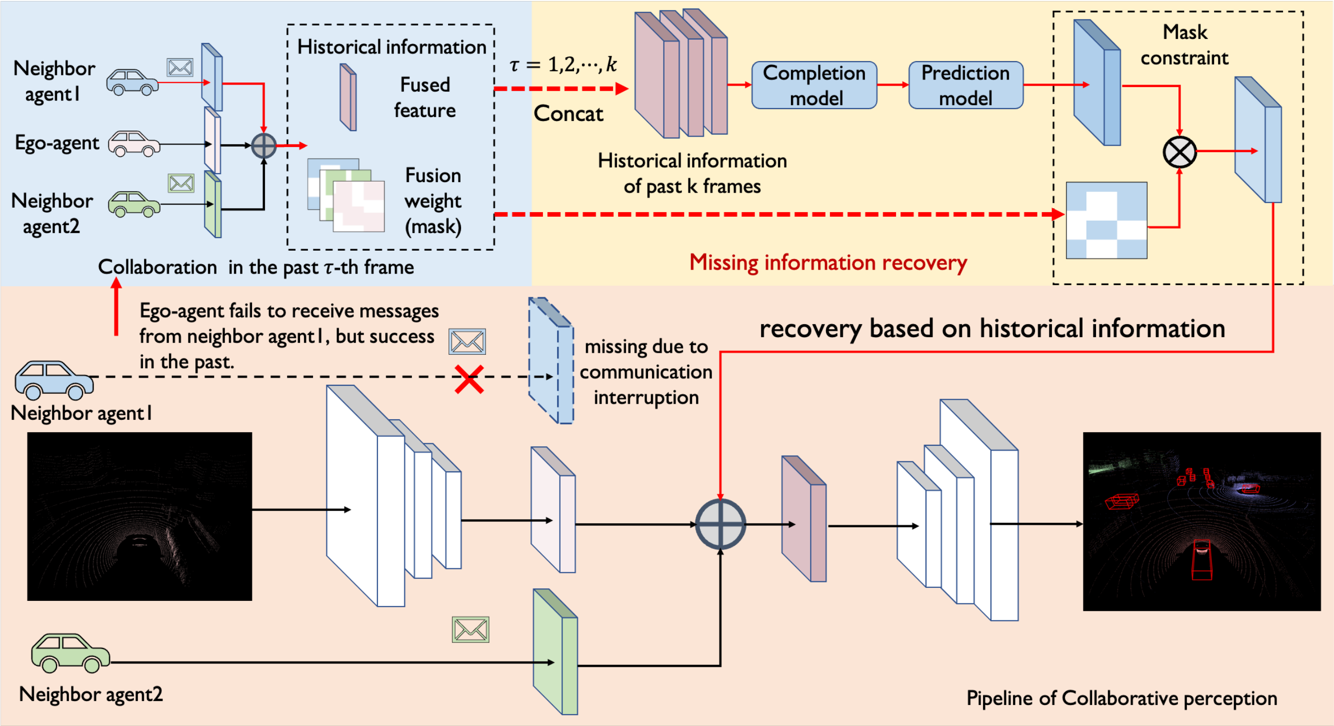

Neighbor agent1
Ego-agent
Neighbor agent2
Historical information
Fused feature
Fusion weight (mask)
Collaboration in the past τ-th frame
τ = 1,2,⋯,k
Concat
Historical information of past k frames
Completion model
Prediction model
Mask constraint
Missing information recovery
Ego-agent fails to receive messages from neighbor agent1, but success in the past.
missing due to communication interruption
recovery based on historical information
Neighbor agent1
Neighbor agent2
Pipeline of Collaborative perception

Figure 2: **Overview of the interruption-aware collaborative perception (IA-RCP) framework.** We recover the missing information from the collaboration history through the missing information recovery process, which is composed of a completion model and a prediction model. The output of the recovery module will be applied to the mask constraint to suppress the noise generated during the recovery. Then we regard it as the feature from a pseudo agent to compensate for the currently disconnected agent.

$$\begin{aligned}\mathbf{Z}_i^{(t-\tau+1\to t)} &= f_{\text{fuse}}\Big(\{\mathbf{F}_{j\to i}^{(t-\tau+1\to t)}\}_{j\in\{i\}\cup\mathcal{R}_i^{(t-\tau+1)}},\\ &\qquad f_{\text{predict}}(\mathbf{Z}_i^{(t-\tau\to t)})\Big),\\ \mathbf{Z}_i^{(t-k\to t)} &= \mathbf{H}_i^{(t-k\to t)},\end{aligned}$$

where $\tau = 1, 2, \cdots, k-1$. f_{predict} is the prediction model to predict the next state of the given features and f_{fuse} is the fusion model to fuse input features. The architecture of f_{predict} and f_{fuse} will be introduced in detail in the following.

After the completion process, we obtain the features that can achieve better consistency. We concatenate the feature after completion $\text{concat}(\{\mathbf{Z}_i^{(t-\tau\to t)}\}_{\tau=1}^{k})$ and feed it into the prediction model.

Prediction model. Since the past feature carries the spatial semantic information at multiple time points, the prediction model f_{predict} leverages a spatial-temporal pyramid network like [Wu *et al.*, 2020] to capture the multi-scale spatial-temporal semantic information. To be specific, we employ two blocks composed of standard 2D convolutions, a degenerated 3D convolution, batch normalization, and ReLU activation. And then, two convolutional layers are used to obtain the output, $\mathbf{F}_{c_i}^{(t)}$, with the size same as the feature to be fused in Eq. 1and c_i is the index of the predicted feature. The prediction model takes the concatenated features after completion and output the recovered missing information $\mathbf{F}_{c_i}^{(t)}$:

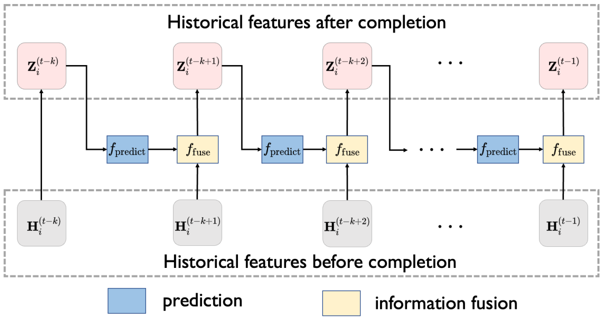


Figure 3: **Completion model.** With the incomplete feature at each historical time stamp, we recurrently predict its state at the next time stamp and fuse it with the features received at next time stamp to obtain the feature after completion.

$$\mathbf{F}_{c_i}^{(t)} = f_{\text{predict}}(\text{concat}(\{\mathbf{Z}_i^{(t-\tau\to t)}\}_{\tau=1}^{k})).$$

Advantages. We conduct the missing information recovery in the feature domain for three reasons. First, it is communication efficient. Compared with sharing raw data, intermediate features are much easier to compress. Recovery in the feature domain can maximized the utility of the shared messages. Second, it makes feature extraction more straightforward. Recovery in the feature domain allow more direct feature extraction and usage because the recovery model is jointly trained with the encoder and decoder. Furthermore, recovery in the feature domain makes the model focus on predicting the most informative features and there is no need to predict the accurate and concrete scenery information, which is more difficult and with less fault tolerance. Third, recovery in the feature domain makes it flexible to apply to multiple tasks and multiple models without changing the design of recovery module.

4.2 Information Fusion

After missing information recovery at time stamp t, agents has the ego feature $\mathbf{F}_i^{(t)}$, the set of the received features from neighbour agents $\{\mathbf{F}_j^{(t)}\}_{j \in \mathcal{R}_i^t}$, and the predicted feature $\mathbf{F}_{c_i}^{(t)}$. We regard $\mathbf{F}_{c_i}^{(t)}$ as the feature from a pseudo agent who can complete the missing information due to the communication interruption at time t. We utilize a spatial attention fusion model to fuse the features from all agents, including the pseudo agent.

Coordinate transformation. Firstly, each agent a_i transforms the features from other agents into its own coordinate system based on their poses, that is,

$$\mathbf{F}_{j \to i}^{(t)} = \xi_{j \to i}(\mathbf{F}_j^{(t)}),$$

where $\xi_{j \to i}$ is the transformation principle based on the poses of the two agents a_i and a_j.After coordinate transformation, all features are supported in the same coordinate systems.

Spatial attention fusion weight calculation. Since multiple agents have distinct locations and views, the features from various agents capture information of different spatial areas. So different spatial cells of the features have different importance to the ego agent. To obtain the most reliable information, we need to give various weights to different spatial cells to strengthen the informative cells and suppress unnecessary and noisy cells during the fusion stage. To calculate the weight, we concatenate each feature with the ego feature along the channel dimension and use multiple 1 × 1 convolutional layers to gradually reduce the number of channel to 1, and finally apply a softmax function at each pixel of all features to get a spatial mask $\mathbf{M}_j^{(t)}$ of each feature $\mathbf{F}_j^{(t)}$, that is,

$$\mathbf{M}_j^{(t)} = \mathrm{softmax}(f_{\mathrm{mask}}(\mathrm{concat}(\mathbf{F}_i^{(t)}, \mathbf{F}_{j \to i}^{(t)}))),$$

for all $j \in \{i, c_i\} \cup \mathcal{R}_i^{(t)}$.

Spatial attention information fusion. With the calculated spatial attention fusion weight, we fuse all features by weighted averaging, that is,

$$\begin{aligned} \mathbf{H}_i^{(t)} =& f_{\mathrm{Fuse}}(\{\mathbf{F}_{j \to i}^{(t)}\}_{j \in \{i, c_i\} \cup \mathcal{R}_i^{(t)}}) \\ =& \sum_{j \in \{i, c_i\} \cup \mathcal{R}_i^{(t)}} \mathbf{M}_j^{(t)} \odot \mathbf{F}_{j \to i}^{(t)}, \end{aligned}$$

where $\odot$ is an element-wise multiplication.

In this way, though some information is missed due to the communication interruption, it can be fused in $\mathbf{H}_i^{(t)}$ if the relevant information has been received in the past. In addition, the spatial attention fusion model could suppress some error and noise produced in the missing information recovery process with the received features. Finally, the fused feature is fed into the decoder to obtain the final perception results.

4.3 Advanced Designs

To tackle some issues in the recovery and collaboration process, we propose two advanced designs to improve the performance of our model.

Spatial attention mask constraint. During the missing information recovery process, some harmful noise or even error may be produced, especially the noise and error contradicting the received information in some certain spatial regions. We aim to eliminate the affect of the recovered information in these regions because the received features are enough for perception.

We suppress the background noise with the help of the spatial attention fusion weight, shown in the right top of Figure 2. If agent a_i lost contact with an agent a_{lose} at time t but has received messages from a_{lose} in history, we transform the latest fusion weight of the feature from agent a_{lose} into the coordinate system of time t and set the element less than a certain threshold to 0 and other elements to 1, then we get the mask $\mathbf{M}_l^*$ of the agent a_{lose}, which can reflect the informative region of the features from a_{lose}. Before fusing the recovered feature $\mathbf{F}_{c_i}^{(t)}$ to compensate the missing feature from a_{lose}, we multiply it by the mask to suppress the information in the regions irrelevant to a_{lose}. Therefore, the potential noise and error in these regions are suppressed together. The final formulation of fused feature at time stamp t can be written as:

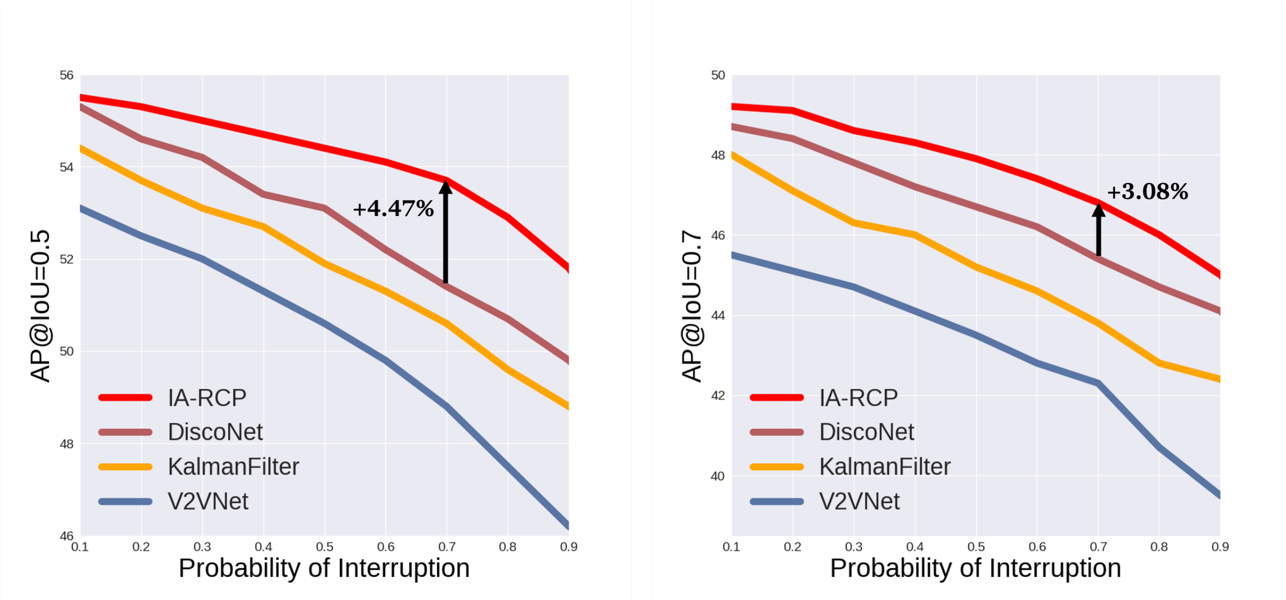
$$\mathbf{H}_i^{(t)} = f_{\mathrm{fuse}}(\{\mathbf{F}_{j \to i}^{(t)}\}_{j \in \{i\} \cup \mathcal{R}_i^{(t)}}, \{\mathbf{F}_{c_i}^{(t)} \odot \mathbf{M}_l^*\}_{l \in \mathcal{N}_i \backslash \mathcal{R}_i^{(t)}}).$$

Curriculum learning. To enable our model to handle different communication conditions, the probability of communication interruption at each iteration is randomly sampled from $(0, 1)$ during the training stage. However, the training loss is closely related to the interruption probability, resulting in the unstable training that will affect the performance.

To tackle this issue, we adopt the idea of curriculum learning [Bengio *et al.*, 2009] that learning first starts with only easy examples of a task and then gradually increases the task difficulty. It is believed that collaborative perception with low interruption is an easier case for our framework's prediction and detection models. Therefore we make the training process begin with low interruption probability, increase the interruption probability range, and gradually add more difficult samples into the training set .

4.4 Loss Function

The proposed IA-RCP framework does not need to be supervised by additional loss function except for the loss function of the perception task itself, which makes it easy to applied and transferred into multiple perception tasks. In this work, we validate the method with the task of collaborative 3D detection based on LIDAR point clouds. The results of detection $\widehat{\mathbf{Y}}$ includes classification result $\widehat{\mathbf{Y}}_{\mathrm{cls}}$ and location result $\widehat{\mathbf{Y}}_{\mathrm{loc}}$. The first one achieves classification to decide the probability of being a target. The second one is a regression problem who calculates the parameters about bounding boxes. As a result, We train our model to minimize a loss function which combines of classification and location loss. The classification loss is a cross-entropy loss calculated over each location



(a) Performance in AP@IOU 0.5

(b) Performance in AP@IOU 0.7

Figure 4: Perception Performance with different interruption probability. Our IA-RCP framework outperforms the baselines significantly.

and pre-defined box. The regression loss is a smooth L1 loss over six variables include central coordinates, length, width and sin and cos value to represent angle of the boxes. $\mathbf{Y}$ is the ground truth of detection which consist of classification ground truth $\mathbf{Y}_{\text{cls}}$ and location ground truth $\mathbf{Y}_{\text{loc}}$. The loss function can be formulated as:

$$\mathrm{L}(\widehat{\mathbf{Y}}, \mathbf{Y}) = \mathrm{L}_{\mathrm{CE}}\left(\widehat{\mathbf{Y}}_{\text{cls}}, \mathbf{Y}_{\text{cls}}\right) + \mathrm{L}_{\text{smooth}-\mathrm{L1}}\left(\widehat{\mathbf{Y}}_{\text{loc}}, \mathbf{Y}_{\text{loc}}\right).$$

5 EXPERIMENTS

5.1 Dataset, Implementation and Evaluation

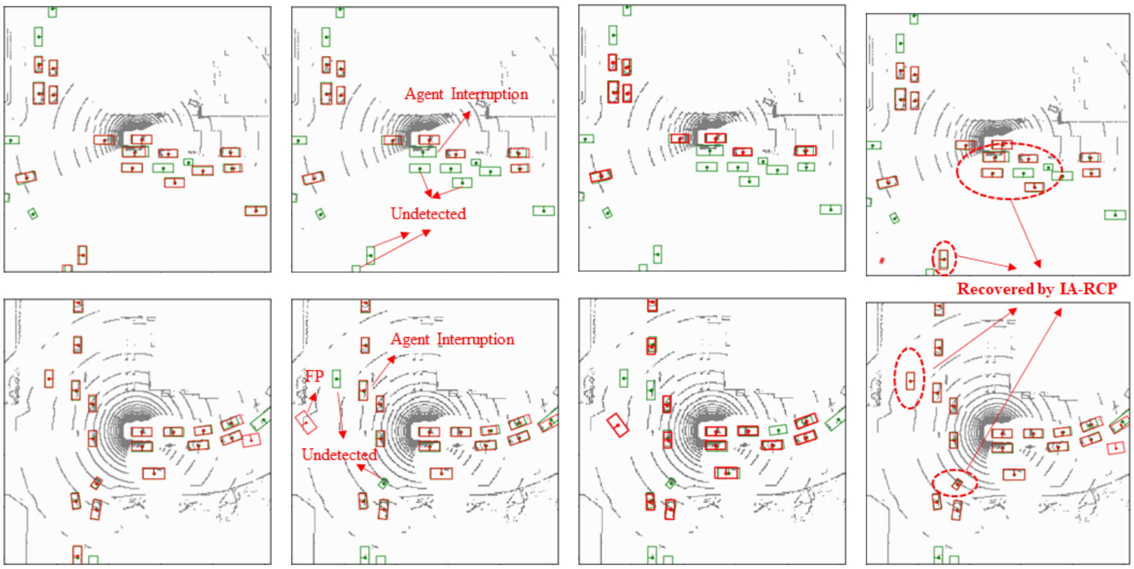
Dataset. We evaluate our method with the task of 3D object detection based on 3D LIDAR point clouds data, which requires detecting the position and size in the 3D space of objects. We employ a public large-scale multi-agent perception dataset, V2X-Sim [Li *et al.*, 2021] to validate our method. V2X-Sim is a simulation dataset built on Carla [Dosovitskiy *et al.*, 2017] and SUMO [Krajzewicz *et al.*, 2012]. The dataset includes 80 scenes in the training set and 11 scenes in the evaluation set. Each scene contains up to five vehicles which can establish communication with each other. The point cloud data is generated with a virtual 32 channels LIDAR with 20Hz rotation frequency, and 5Hz recorded frequency. The max detection range of the LIDAR is 70m, and the number of points per second is up to 250000. We sample 10 frames from each scene in both the training and evaluation sets.

Implementation and evaluation. We employ the architecture of the encoder and decoder of [Li *et al.*, 2021]. The number of the history features k is set to 3. We train our models for 100 epochs using Adam [Kingma and Ba, 2015] optimizer on NVIDIA GeForce RTX 3090 GPU. We set the batch size to 4 and the initial learning rate to 0.001, which will decay to 0.0005 after the 50th epoch. We adopt the generic detection evaluation metric: Average Precision (AP) at Intersection-over-Union (IoU) threshold of 0.5 and 0.7. We test the model on the test set at nine different interruption probabilities ranging from 0.1 to 0.9.

5.2 Baselines

Existing collaborative perception strategies. Since the proposed IA-RCP framework is the first to address the communication interruption issue in collaborative perception, we take existing collaborative perception strategies as the baselines. i) *V2VNet* [Wang *et al.*, 2020]: a state-of-the-art collaboration strategy, which fuses received features and the ego feature with a graph neural network; ii) *DiscoNet* [Li *et al.*, 2021]: a state-of-the-art collaborative perception method, which fuses features with a collaboration graph distilled by a teacher model employing early data fusion. We apply IA-RCP framework to the two baselines to validate the effectiveness of our method.

Prediction from history results by Kalman Filter. To illustrate the effectiveness of the prediction in the feature domain in IA-RCP, we compare it with the prediction from Kalman Filter based on history. Given the bounding boxes detected with the incomplete features by DiscoNet, we implemented a naive Kalman Filter to track each object and estimate its future state. We used the Hungarian algorithm to match prediction and measurement in the tracking process. Finally, we combined the predicted boxes with the detection results based on currently received features with a simple selection. This baseline method is named by *KalmanFilter*.



(a) DiscoNet (without interruption)
(b) DiscoNet (with interruption)
(c) Kalman Filter (with interruption)
(d) IA-RCP (with interruption)

Figure 5: IA-RCP qualitatively improves the perception performance with communication interruption. Each row represents an example and each column represents a method: (a) the detection results of *DiscoNet* [Li *et al.*, 2021] without any communication interruption; (b) the detection results of DiscoNet when interruption occurs; (c) the detection results of *KalmanFilter*; (d) the detection results of IA-RCP. Our method recovers most of the boxes which are lost due to the communication interruption.

5.3 Results

Figure 4 shows the collaborative perception performance in term of AP(@IOU = 0.5/0.7) at different interruption probability from 0.1 to 0.9. We see that: i) The performance of proposed IA-RCP framework outperforms the two baselines, especially in the case with high interruption probability. IA-RCP outperforms *DiscoNet* by 4.47% in AP@0.5 and 3.08% AP@0.7 at interruption probability = 0.7. ii) The prediction from the perception results by *Kalman Filter* does not work well, and its performance is even worse than the baseline in most situations, suggesting that it is hard to obtain good recovery from the output of history.

The results validate the effectiveness of the proposed IA-RCP framework, which can alleviate the communication interruption issue in collaborative perception. The results suggests that compared with recovery in the output domain, recovery in the feature domain is the more reliable choice, which can take advantage of the history information.

5.4 Ablation Study

Table 1 shows the ablation study of the two advanced designs, including spatial attention mask constraint and curriculum learning. Vinilla IA-RCP denotes the framework without the two designs. We see that: each design can improve the performance of the framework and improve more when both are integrated together, which demonstrates the effectiveness of the two designs.

Table 1: Ablation study of the effect of the two advanced designs. **Curri** represents the curriculum learning and **Mask** represents spatial attention mask constraint. Mean AP is calculated by averaging the performance result with the nine probability $(0.1 \sim 0.9)$.

Method	mean AP @IOU=0.5	mean AP @IOU=0.7
vanilla **IA-RCP**	53.5	47.1
+ **Curri**	53.7	47.3
+ **Mask**	54.0	47.4
+ **Curri** & **Mask**	**54.2**	**47.7**

5.5 Visualization of Qualitative Results

Fig. 5 shows the comparison of different methods in case of communication with/without interruption. Each row represents a sample and each column represents the same method and scenario. Fig.5(a) and 5(b) show the detection results of *DiscoNet* without and with communication interruption respectively. We see that: *DiscoNet* fails to detect some objects with the collaboration with other agents due to communication interruption. Fig.5(c) shows the detection results of *KalmanFilter*, whose prediction is not precise enough to improve performance. Fig.5(d) show the results of our IA-RCP framework. Our method recovers most of the lost boxes due to the communication interruption.

6 CONCLUSION

In this work, we study the problem of collaborative perception with stochastic communication interruption and propose an interruption-aware robust collaborative perception (IA-RCP) framework. Its core idea is to utilize historical information to recover missing information due to communication interruption. In addition, We propose two further designs to improve the quality of missing information recovery: spatial attention mask for background suppression, and curriculum learning strategy for more stable training. Comprehensive results and quantitative results show that the proposed IA-RCP framework brings significant benefits to alleviate the effect of communication interruption.

References

[Alotaibi *et al.*, 2019] Ebtehal Turki Alotaibi, Shahad Saleh Alqefari, and Anis Koubaa. Lsar: Multi-uav collaboration for search and rescue missions. *IEEE Access*, 7:55817–55832, 2019.

[Araniti *et al.*, 2013] Giuseppe Araniti, Claudia Campolo, Massimo Condoluci, Antonio Iera, and Antonella Molinaro. Lte for vehicular networking: a survey. *IEEE Communications Magazine*, 51:148–157, 2013.

[Arnold *et al.*, 2020] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[Chen *et al.*, 2018] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2018.

[Chen *et al.*, 2019] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019.

[Dosovitskiy *et al.*, 2017] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[Jiang and Delgrossi, 2008] Daniel Jiang and Luca Delgrossi. Ieee 802.11p: Towards an international standard for wireless access in vehicular environments. *VTC Spring 2008 - IEEE Vehicular Technology Conference*, pages 2036–2040, 2008.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[Krajzewicz *et al.*, 2012] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4), 2012.

[Lei *et al.*, 2016] Lei Lei, Yiru Kuang, Nan Cheng, Xuemin Shen, Zhangdui Zhong, and Chuang Lin. Delay-optimal dynamic mode selection and resource allocation in device-to-device communications—part i: Optimal policy. *IEEE Transactions on Vehicular Technology*, 65:3474–3490, 2016.

[Li *et al.*, 2020] Zhi Li, Ali Vatankhah Barenji, Jiazhi Jiang, Ray Y Zhong, and Gangyan Xu. A mechanism for scheduling multi robot intelligent warehouse system face with dynamic demand. *Journal of Intelligent Manufacturing*, 31(2):469–480, 2020.

[Li *et al.*, 2021] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[Liu *et al.*, 2020a] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2020.

[Liu *et al.*, 2020b] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883, 2020.

[Mei *et al.*, 2018] Jie Mei, Kan Zheng, Long Zhao, Yong Teng, and Xianbin Wang. A latency and reliability guaranteed resource allocation scheme for lte v2v communication systems. *IEEE Transactions on Wireless Communications*, 17:3850–3860, 2018.

[Miller *et al.*, 2020] Aaron Miller, Kyungzun Rim, Parth Chopra, Paritosh Kelkar, and Maxim Likhachev. Cooperative perception and localization for cooperative driving. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1256–1262. IEEE, 2020.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[Scherer *et al.*, 2015] Jürgen Scherer, Saeed Yahyanejad, Samira Hayat, Evsen Yanmaz, Torsten Andre, Asif Khan, Vladimir Vukadinovic, Christian Bettstetter, Hermann Hellwagner, and Bernhard Rinner. An autonomous multi-uav system for search and rescue. In *Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*, pages 33–38, 2015.

[Shi *et al.*, 2019] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019.

[Wang *et al.*, 2020] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and

Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 605–621, 2020.

[Wu *et al.*, 2020] Pengxiang Wu, Siheng Chen, and Dimitris N. Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11382–11392, 2020.

[Zaccaria *et al.*, 2021] Michela Zaccaria, Mikhail Giorgini, Riccardo Monica, and Jacopo Aleotti. Multi-robot multiple camera people detection and tracking in automated warehouses. In *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)*, pages 1–6. IEEE, 2021.