Bandwidth-Adaptive Spatiotemporal Correspondence Identification for Collaborative Perception

Peng Gao[1], Williard Joshua Jose[2], Hao Zhang[2]

***Abstract*— Correspondence identification (CoID) is an essential capability for multi-robot collaborative perception, which allows a group of robots to consistently refer to the same objects in their own fields of view. In real-world applications, such as connected autonomous driving, connected vehicles cannot directly share their raw observations due to the limited communication bandwidth. To address this challenge, we propose a novel approach of bandwidth-adaptive spatiotemporal CoID for collaborative perception, where robots interactively select partial spatiotemporal observations to share with others, while adapting to the communication constraint that dynamically changes over time. We evaluate our approach over various scenarios in connected autonomous driving simulations. Experimental results have demonstrated that our approach enables CoID and adapts to the dynamic change of bandwidth constraints. In addition, our approach achieves 8%-56% overall improvements in terms of covisible object retrieval for CoID and data sharing efficiency, which outperforms the previous techniques and achieves the state-of-the-art performance.**

I. Introduction

Multi-robot systems have earned significant attention over the past few decades, primarily owing to their proven reliability and efficiency in tackling cooperative tasks. These tasks encompass a broad spectrum, including collaborative manufacturing, multi-robot search and rescue, and connected autonomous driving. To facilitate effective collaboration among robots, a fundamental capability is collaborative perception. This capability allows multiple robots to share perceptual data about their surrounding environment, thereby fostering a shared situational awareness.

As a key component of collaborative perception, correspondence identification (CoID) plays a critical role, with the goal of identifying the same objects concurrently observed by multiple robots within their respective field of view. As shown in Figure 1, when a pair of connected vehicles meet at a street intersection, it is important for these vehicles to accurately identify the correspondence of street objects observed in their observations. Given the identified correspondences, connected vehicles can effectively refer to the same objects and estimate their relative poses among robots, thus facilitating further collaboration.

Given the importance of CoID, a variety of techniques are developed, which fall into two primary categories: learning-free and learning-based approaches. Learning-free techniques comprise keypoint-based visual association, geometric-based spatial matching, and synchronization techniques. In contrast,

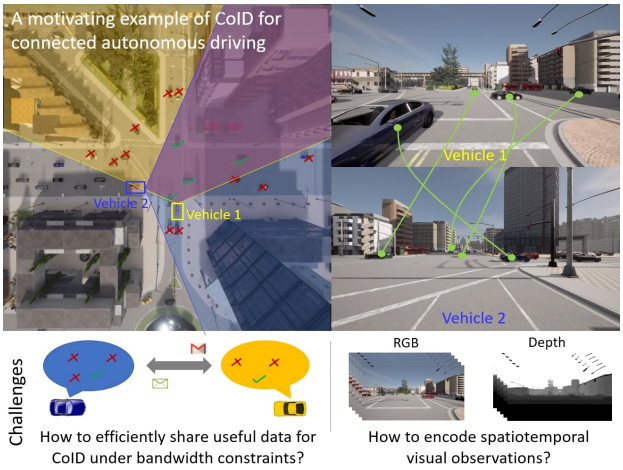


Fig. 1. A motivating example of CoID under the communication bandwidth constraint for collaborative perception in connected autonomous driving. In order to enable connected vehicles to refer to the same street objects, they must efficiently share spatiotemporal information to identify object correspondences, while satisfying the communication bandwidth constraint.

learning-based methods harness deep learning, employing convolutional neural networks (CNNs) for object re-identification from different perspectives and graph neural networks (GNNs) for deep graph matching. Although the previous methods show promising performance, two significant challenges have not been well addressed yet. The first challenge is caused by the limited communication bandwidth for data sharing in multi-robot systems. In the real-world setting, the maximum bandwidth designated for vehicle-to-everything (V2X) communication is around 7.2 Mbps [1]. It is unrealistic to assume that the available bandwidth allows for the sharing of raw observations, particularly in scenarios involving crowded environments. The second challenge is caused by the need to share temporal data among connected robots, which further demands a much larger bandwidth compared to sharing single-frame observations.

In order to address the challenges, we propose a novel bandwidth-adaptive method to perform spatiotemporal CoID. We develop a spatiotemporal graph representation to encode spatiotemporal visual information of street objects observed by each vehicle. Each node encodes a street object. Each spatial edge encodes the spatial relationship of a pair of objects, and each temporal edge is designed to track each object's motion. Given this graph, our approach formulates CoID as a progressive graph-matching problem, while adapting data sharing to the dynamically changing bandwidth.

[1]North Carolina State University, email: pgao5@ncsu.edu
[2]University of Massachusetts Amherst, email: {wjose, hao.zhang}@umass.edu

Specifically, we develop a heterogeneous attention network that generates node features by integrating the objects' visual, spatial, and temporal cues. We develop a pooling operation that explicitly encodes the spatial and temporal cues to generate a graph-level embedding to encode the global scene. Then, we introduce our new framework to enforce individual robots to progressively share nodes that are most likely to be also observed by their collaborating vehicles to maximize the CoID performance while satisfying the given communication bandwidth constraints.

The key contribution of this paper is the introduction of the novel bandwidth-adaptive CoID approach for collaborative perception. Specific novelties include:

- A novel progressive CoID framework that allows adaptive data sharing according to current communication bandwidth, which enables connected vehicles to fully utilize the allocated bandwidth under the communication constraints.
- A new heterogeneous attention network to integrate visual, spatial, and temporal cues of objects in a unified way, which encodes both current and historical cues to enhance the expressiveness of node features for CoID.
- A new heterogeneous graph pooling operation to generate a graph-level embedding of the comprehensive scene, which explicitly encodes the importance of the temporal and spatial cues, as well as compresses the spatiotemporal observations.

II. Related Work

A. Connected Autonomous Driving

The growing interest in connected autonomous driving, driven by collaborative perception among connected agents, has led to various research efforts. Methods are generally classified into raw-based early collaboration, output-based late collaboration, and feature-based intermediate collaboration. Early collaboration fuses raw sensor data from connected agents onboard for vision tasks [2], while late collaboration merges multi-agent perception outputs using techniques like Non-Maximum Suppression [3] and refined matching to ensure pose consistency [4]. Intermediate collaboration strikes a balance by sharing compressed features, with methods such as when2com [5], who2com [6], and where2com [7]. Data fusion strategies include concatenation [8], re-weighted summation [9], graph learning [10, 11], and attention-based fusion [12, 13]. Applications span object detection [14], tracking [15], segmentation [16], localization [17], and depth estimation [18]. However, none of these methods adapts to bandwidth limitations, which often prevent the sharing of holistic information.

B. Correspondence Identification

Correspondence Identification (CoID) methods fall into learning-free and learning-based categories. Learning-free approaches include visual appearance techniques like SIFT [19], ORB [20], HOG [21], and TransReID [22], as well as spatial techniques like ICP [23], template matching [24], and graph matching [25, 26]. Synchronization algorithms also contribute through circle consistency enforcement [27] and convex optimization [28]. Learning-based methods primarily use CNNs [29–31] and GNNs [32–34], with hybrid approaches like Bayesian CoID [35, 36] enhancing robustness. However, existing methods struggle to integrate temporal cues, as sharing sequences of frames is constrained by real-world bandwidth limitations. We propose a novel method that integrates visual, spatial, and temporal cues for CoID in a bandwidth-adaptive way.

III. Approach

We discuss our proposed bandwidth-adaptive spatiotemporal CoID method in this section, which is illustrated in Figure 2. We assume that each of the ego vehicle and collaborative vehicle is equipped with a RGB-D camera or a LiDAR sensor. Formally, each vehicle obtains a sequence of observations $\mathcal{O} = \{obs^t, obs^{t+1}, \ldots, obs^{t+T}\}$. Each observation recorded at time t consists of detected objects $obs^t = \{\mathbf{v}_1^t, \mathbf{v}_2^t, \ldots, \mathbf{v}_m^t\}$ where $\mathbf{v}_i \in \mathcal{R}^3$ denotes the attributes of the i-th object detected at time t. Given a sequence of observations $\mathcal{O}$, we represent it as a spatiotemporal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{spat}, \mathcal{E}^{temp}\}$. $\mathcal{V}$ denotes the node set, which contains all the attributes of objects detected in the observation sequence $\mathcal{O}$. $\mathcal{E}^{spat} = \{e_{p,q}^{spat}\}$ denotes the spatial relationships between a pair of objects. $e_{p,q}^{spat}$ denotes the distance between the p-th object and the q-th object recorded at the same time t, otherwise 0. $\mathcal{E}^{temp} = \{e_{p,q}^{temp}\}$ denotes the temporal relationships of the same object recorded at different times. If $\mathbf{v}_p^{t_1}$ and $\mathbf{v}_q^{t_2}$ are the same object recorded at time t_1 and t_2, then $e_{p,q}^{temp} = t_2 - t_1$, otherwise 0.

A. Spatiotemporal Graph Embedding

Given the spatiotemporal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{spat}, \mathcal{E}^{temp}\}$, we introduce a new heterogeneous graph attention network that encodes objects' visual, spatial and temporal cues for node-level embedding. Formally, the node embedding vectors are defined as $\{\mathbf{m}_i\}^N = \psi(\mathcal{G})$, where ψ is the heterogeneous attention network. Specifically, we first project each node feature to the same feature space as follows:

$$\mathbf{h}_i = \mathbf{W}_v \mathbf{v}_i \tag{1}$$

where $\mathbf{h}_i$ denotes the projected feature of the i-th node, $\mathbf{W}_v$ denotes the associating weight matrix, and $\mathbf{v}_i$ denotes the original feature vector of the i-th object. Then we compute the self-attention of each node given different types of edges, defined as follows:

$$\alpha_{i,j} = \frac{\exp\left(\sigma([\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_j || \mathbf{W}_e e_{i,j}])\right)}{\sum_{k \in \mathcal{N}^{\Psi}(i)} \exp(\sigma([\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_k || \mathbf{W}_e e_{i,k}]))} \tag{2}$$

where $\alpha_{i,j}$ is the attention from node j to node i, σ denotes the ReLu activation function, $||$ denotes the concatenation operation, $\mathbf{W}$ and $\mathbf{W}_e$ are weight matrices. This attention $\alpha_{i,j}$ is obtained by comparing the centered i-th node with its neighborhood nodes. Then, we normalize the attention using the SoftMax function. To encode spatial and temporal relationships of objects, we add edge attributes into the learning process given the edge connection $\mathcal{N}^{\Psi}(i)$, where $\Psi \in \{spat, temp\}$ denotes two types of edges connected to

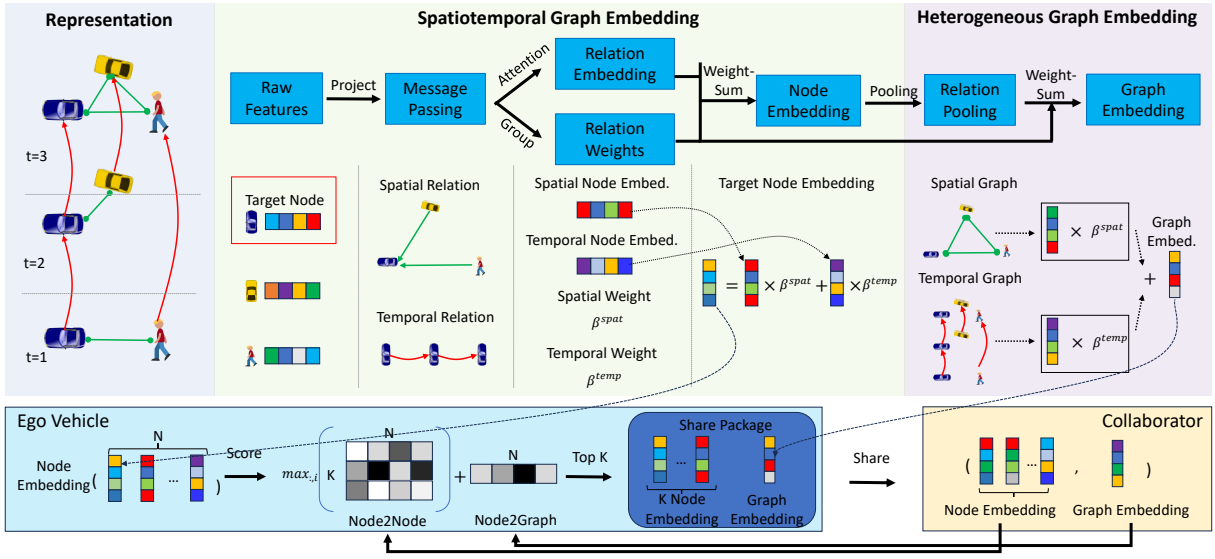


Fig. 2. An overview of our proposed bandwidth-adaptive spatiotemporal CoID approach. A sequence of observations is represented as a spatiotemporal graph. A spatiotemporal graph attention network is used to generate node-level embeddings by integrating spatiotemporal visual cues. Then, a heterogeneous graph pooling operation is designed to produce comprehensive graph-level embeddings that explicitly encode the importance of spatial and temporal cues. Leveraging both node-level and graph-level embeddings, our approach enables the sharing of node candidates that are likely to be observed by collaborators, resulting in effective data sharing that adapts to communication bandwidth constraints.

the i-th node. Given these two types of edges, we get two attention values $\alpha_{i,j}^{spat}, \alpha_{i,j}^{temp}$. Then, we compute the node embedding vector as:

$$\mathbf{h}_i = \sigma\left(\mathbf{W}\mathbf{h}_i + \sum_{j \in \mathcal{N}^{\Psi}(i)} \alpha_{i,j}(\mathbf{W}\mathbf{h}_j + \mathbf{W}_e e_{i,j})\right) \quad (3)$$

For the same node i, we compute its spatial and temporal embedding vectors $\mathbf{h}_i^{spat}$ and $\mathbf{h}_i^{temp}$ given its associated attentions $\alpha_{i,j}^{spat}, \alpha_{i,j}^{temp}$. They are computed via aggregating the node and edge embedding features weighted by attention values. We also use a multi-head mechanism to enable the network to catch a richer representation of the embedding. Multi-head embedding vectors are concatenated after intermediate attention layers.

To combine these two spatial and temporal embedding vectors of the same node, we learn the weights of spatial and temporal relationships of nodes to indicate their importance, which is defined as follows:

$$\beta^{spat} = \frac{1}{|\mathcal{V}^{spat}|} \sum_{i \in \mathcal{V}^{spat}} \mathbf{q}^T \tanh(\mathbf{W}_b \mathbf{h}_i + \mathbf{b}) \quad (4)$$

$$\beta^{temp} = \frac{1}{|\mathcal{V}^{temp}|} \sum_{i \in \mathcal{V}^{temp}} \mathbf{q}^T \tanh(\mathbf{W}_b \mathbf{h}_i + \mathbf{b}) \quad (5)$$

where $\mathbf{W}_b$ denotes the weight matrix, $\mathbf{b}$ is the bias vector, $\mathbf{q}$ denotes the learnable edge-specific attention vector. The learnable parameters are shared for all spatial and temporal relationships. Then, the spatial and temporal weights are normalized through SoftMax, defined as follows:

$$\beta^{spat} = \frac{\exp(\beta^{spat})}{\exp(\beta^{spat} + \beta^{temp})} \quad (6)$$

$$\beta^{temp} = \frac{\exp(\beta^{temp})}{\exp(\beta^{spat} + \beta^{temp})} \quad (7)$$

where $\beta^{spat}, \beta^{temp}$ denote the weights of spatial and temporal relationships for CoID. The higher the value, the larger the importance of the type of relationship. The final node embedding is computed as:

$$\mathbf{m}_i = \sum_{\Psi \in \{spat, temp\}} \beta^{\Psi} \mathbf{h}_i^{\Psi} \quad (8)$$

where $\mathbf{m}_i$ denotes the final embedding vector of the i-th object in the spatiotemporal graph by integrating object positions, and spatiotemporal relationships.

B. Heterogeneous Graph Pooling

Due to the communication bandwidth constraint, the collaborator robot can only share partial nodes with the ego robot. To also share the comprehensive information of the collaborator's observations for CoID, we further propose a novel graph pooling operation, which compresses the collaborator's spatiotemporal graph into a single vector and integrates with the proposed heterogeneous graph network in a principled way.

Specifically, we decompose the graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{spat}, \mathcal{E}^{temp}\}$ into two separate graphs, including $\mathcal{G}^{spat} = \{\mathcal{V}, \mathcal{E}^{spat}\}$ and $\mathcal{G}^{temp} = \{\mathcal{V}, \mathcal{E}^{temp}\}$. Then, we perform ASAPooling [37] on both of the spatial and temporal graphs, which is defined as follows:

$$\mathbf{z}^{\Psi} = \phi(\{\mathbf{m}_i\}^N, \mathcal{E}^{\Psi}), \Psi = \{spat, temp\} \quad (9)$$

where ϕ is the ASAP pooling operation, and $\mathbf{z}^{\Psi} = \{\mathbf{z}^{spat}, \mathbf{z}^{temp}\}$ is the graph-level embedding concerning the spatial and temporal edges. By taking advantage of the

weights of spatiotemporal relationships, we compute the final graph embedding vector as:

$$\mathbf{z} = \beta^{spat}\mathbf{z}^{spat} + \beta^{temp}\mathbf{z}^{temp} \tag{10}$$

where $\mathbf{z}$ denotes the graph embedding vector of the spatiotemporal graph $\mathcal{G}$, which is computed by the sum of spatial and temporal embedding vectors weighted by the importance of different types of relationships.

C. Bandwidth-Adaptive CoID

Given the node and graph-level embedding vectors generated by each robot, we design a novel interactive approach to perform CoID. Specifically, on the ego robot side, we have ego node embedding vectors $\{\mathbf{m}\}^N$, collaborator node embedding vectors $\{\mathbf{m}'\}^K$ and the graph embedding vector $\mathbf{z}'$. K denotes the communication bandwidth that allows the maximum number of nodes to share. Then we compute matching scores to indicate the probabilities of ego nodes to appear in the collaborator's observations. The computation is defined as:

$$\begin{aligned} \mathbf{S}_{i,j}^{node} &= \exp(-||\mathbf{m}_i - \mathbf{m}'_j||_2) \\ \mathbf{s}_i^{graph} &= \exp(-||\mathbf{m}_i - \mathbf{z}'||_2) \end{aligned} \tag{11}$$

where $\exp()$ denotes the exponential operator, $\mathbf{S}^{node} \in \mathbb{R}^{N\times K}$ represents the similarity between pairs of ego-collaborator node embedding vectors and $\mathbf{s}^{graph} \in \mathbb{R}^N$ denotes the similarity between the ego node embedding vector and collaborator graph embedding vector. Finally, we select the Top_K candidates given $\{\mathbf{S}_{i,j}^{node}\}$ and $\{\mathbf{s}_i^{graph}\}$. Specifically,

$$\mathbf{s} = \lambda\mathbf{s}^{graph} + (1-\lambda)\max_{:,i}\mathbf{S}^{node}, \quad i = 1, 2, \ldots, N \tag{12}$$

where $\mathbf{s}$ denotes the final matching score to select correspondence candidates, λ denotes a hyperparameter to indicate the importance of node2node similarity and node2graph similarity, and $\max_{:,i}\mathbf{S}^{node}$ denotes the maximum elements in each column of $\mathbf{S}^{node}$. Then, the top-K candidates are selected as $\{\mathbf{m}_i\}^K = Top_K(\mathbf{s})$ and share with the collaborator. This process interactively runs on both ego and collaborator vehicles until reaching the maximum interaction number.

Given the received candidate set $\mathcal{M} = \{\mathbf{m}'\}^m$, where m is the total number of nodes sent from the collaborator and the ego robot's graph $\mathcal{G}$, we perform graph matching to identify the correspondences of objects detected in multi-robot observations. Specifically, we compute the similarity $\mathbf{A} \in \mathbb{R}^{n\times m}$ by:

$$\mathbf{A}_{i,j} = \mathbf{m}_i^\top \mathbf{m}'_j \tag{13}$$

where the similarity score $\mathbf{A}$ contains similarities of spatiotemporal visual features of the objects, $\mathbf{m}_i \in \mathcal{G}$ denotes the node embedding vectors in the ego graph $\mathcal{G}$, and $\mathbf{m}'_j \in \mathcal{M}$ denotes the node embedding vectors shared by the collaborator. The correspondences can be identified as $\mathbf{Y} = \text{SoftMax}(\mathbf{A})$, where $\mathbf{Y}$ is the correspondence matrix with $\mathbf{Y}_{i,j} = 1$ denoting the correspondence between the i-th object in $\mathcal{G}$ and the j-th object in $\mathcal{M}$, otherwise $\mathbf{Y}_{i,j} = 0$. We use the circle loss to train our network, which is defined as:

$$\begin{aligned} L_{\mathcal{G}'\to\mathcal{G}}(\mathbf{D}) = \sum_{\mathbf{v}'_i\in\mathcal{V}'} \log\Bigg(1 + \sum_{\mathbf{v}_j\in\mathcal{V}^p} \exp\left[\gamma(\mathbf{D}_{i,j} - \delta_p)^2\right] \\ + \sum_{\mathbf{v}_k\in\mathcal{V}^n} \exp\left[\gamma(\delta_n - \mathbf{D}_{i,k})^2\right]\Bigg) \end{aligned} \tag{14}$$

$L_{\mathcal{G}'\to\mathcal{G}}$ describes the loss given node sets $\mathcal{V} = \{\mathcal{V}^p, \mathcal{V}^n\}$ and $\mathcal{V}'$. $\mathcal{V}^p$ denotes the positive nodes that have corresponding nodes in graph $\mathcal{G}'$. $\mathcal{V}^n$ denotes the negative nodes that have no corresponding nodes in graph $\mathcal{G}'$. δ_p and δ_n are two hyperparameters, which denote the positive and negative margins separately. γ denotes the scale factor. $\mathbf{D} = \{D_{i,j}\}^{n\times n'}$ denotes the distance matrix. We design two different distance matrices in this paper, including $\mathbf{D}_{i,j}^{node} = ||\mathbf{m}_i - \mathbf{m}'_j||_2$ denoting the distance between a pair of node feature vectors, and $\mathbf{D}_i^{graph} = ||\mathbf{m}_i - \mathbf{z}||_2$ denoting the distance between the node feature and graph feature. Similar to the definition of $L_{\mathcal{G}'\to\mathcal{G}}$, we can compute the loss $L_{\mathcal{G}\to\mathcal{G}'}$ given node sets $\mathcal{V}$ and $\mathcal{V}' = \{\mathcal{V}'^p, \mathcal{V}'^n\}$. The final loss is:

$$\begin{aligned} L = \frac{1}{4}\big(L_{\mathcal{G}\to\mathcal{G}'}(\mathbf{D}^{node}) + L_{\mathcal{G}'\to\mathcal{G}}(\mathbf{D}^{node}) \\ + L_{\mathcal{G}\to\mathcal{G}'}(\mathbf{D}^{graph}) + L_{\mathcal{G}'\to\mathcal{G}}(\mathbf{D}^{graph})\big) \end{aligned} \tag{15}$$

The full algorithm is presented in the Appendix.

IV. Experiment

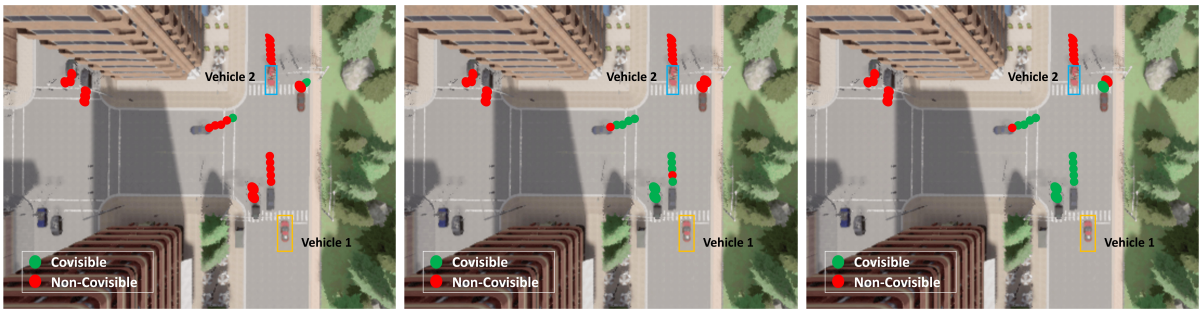
A. Experimental Setups

We have developed a high-fidelity connected autonomous driving (CAD) simulator by seamlessly integrating two open-source platforms: CARLA [38] and SUMO [39]. In the simulations, we collect data to train and evaluate our proposed approach. The details are in the Appendix.

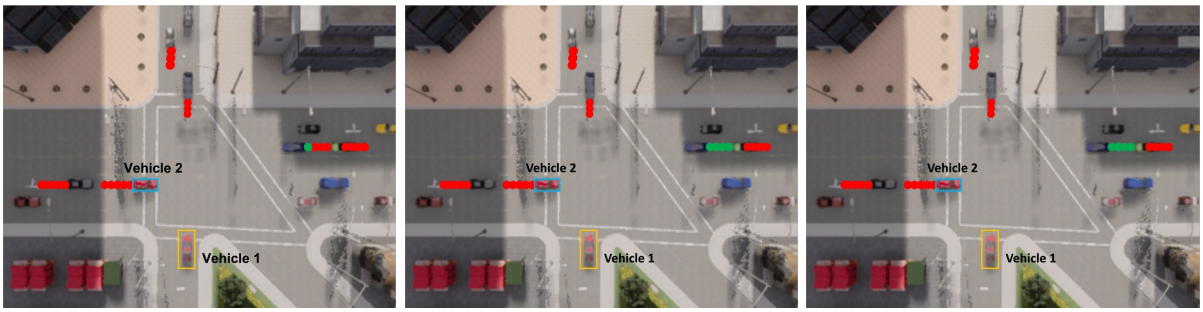
We implement the full version of our method for bandwidth-adaptive spatiotemporal CoID, which uses $Top_K(\mathbf{s})$ to select candidates for CoID. In addition, we implement a baseline method labeled as **Ours-NE**, which uses $Top_K(\mathbf{S}^{node})$ to only compare node embeddings (NE) of the collaborators and the ego vehicle to select candidates. Furthermore, we compare our approach with three previous CoID methods, including:

- Graph convolutional neural network for graph matching (**GCN-GM**) that uses the spline kernel to aggregate visual-spatial information of objects for CoID [40].
- Deep graph matching consensus (**DGMC**) that performs an iterative refinement process on the similarity matrix given the consensus principle [41].
- Bayesian deep graph matching (**BDGM**) that performs CoID under a Bayesian framework and quantifies correspondences' uncertainties to reduce non-covisible objects [35].

These comparison methods use randomly selected query nodes sent from collaborators. None of the comparison methods are capable of integrating temporal information and addressing communication bandwidth adaptation issues.



(a) DGMC (b) Ours (c) GroundTruth



(d) DGMC (e) Ours (f) GroundTruth

Fig. 3. Qualitative results obtained by our approach in normal traffic scenarios (the first row) and crowd traffic scenarios (the second row), as well as comparisons with the DGMC method and the ground truth. The sequence of points denotes a history of observations, which consists of 1-5 points indicating object locations in the past 1-5 time steps. Red points denote non-covisible objects that can only be observed by Vehicle 1. Green points represent the identified covisible objects that can be observed by both Vehicle 1 (in the orange bounding box) and Vehicle 2 (in the blue bounding box). All results are computed in the setup that Vehicle 1 receives information from Vehicle 2, and Vehicle 2 aims to share covisible objects with Vehicle 1.

TABLE I

QUANTITATIVE RESULTS IN THE CAD SIMULATIONS. THE BIS METRIC IS USED TO EVALUATE RECALL AND COMMUNICATION EFFICIENCY OF COID.

Scenario	1 (Normal)			2 (Normal)			3 (Normal)			4 (Crowd)		
Method	Precision ↑	Recall ↑	BIS ↑	Precision ↑	Recall ↑	BIS ↑	Precision ↑	Recall ↑	BIS ↑	Precision ↑	Recall ↑	BIS ↑
GCN-GM [40]	0.5457	0.5294	1.1056	0.3771	0.7074	1.6801	0.5441	0.5373	1.4359	0.3923	0.5724	1.9737
BDGM [35]	0.5420	0.5345	1.1222	0.3538	0.6806	1.6163	0.5488	0.5296	1.4153	0.3852	0.5624	1.9390
DGMC [34]	0.5588	0.5518	1.1331	0.3665	0.7052	1.6747	0.5544	0.5397	1.4421	0.3896	0.5716	1.9709
Ours-NE	0.5995	0.6968	1.3993	0.4867	0.8851	2.4345	0.6478	0.8890	1.5290	0.4503	0.6166	2.2979
Ours	**0.6102**	**0.7123**	**1.4305**	**0.5044**	**0.9241**	**2.5418**	**0.6765**	**0.9261**	**1.5291**	**0.4737**	**0.6756**	**2.5008**

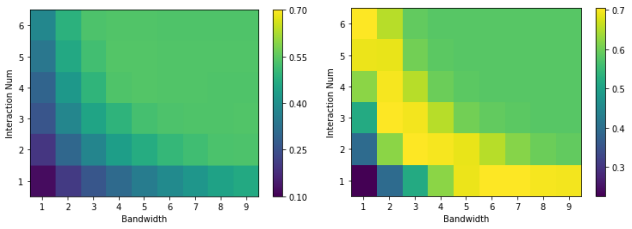
To quantitatively evaluate the CoID performance, we employ the following metrics:

- **Precision** is defined as the ratio of the retrieved objects with correspondences over all retrieved correspondences.
- **Recall** is defined as the ratio of the retrieved objects with correspondences over the ground truth correspondences.
- **F1 Score** is to evaluate the overall performance of CoID methods, which is defined as: $\frac{2\times Precision\times Recall}{Precision+Recall}$
- **BIS**, following the recent work [6], it is defined as: $Recall * \frac{\#Query}{\#Interact*\#Shared}$, where $\#Query$ is the number of query nodes (collaborator's observed nodes), $\#Interact$ denotes the times of interaction, and $\#Shared$ denotes the number of shared nodes in each interaction. The second term describes the ratio of shared nodes compared with all query nodes. Thus, a higher BIS value indicates a higher recall and a lower amount of shared data. BIS $= 1$ means that the collaborator shares all of its observations with the ego vehicle and the recall is 1.

B. Results in Normal Traffic Scenarios

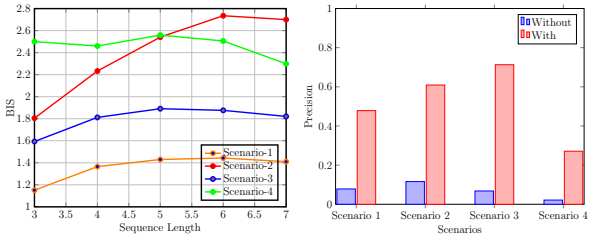
We designed three normal traffic scenarios (Scenarios 1, 2, and 3) in the CAD simulation to evaluate our approach. These scenarios cover key challenges like complex street object interactions, occlusion, limited communication bandwidth, missing objects in observation sequences, and dynamic traffic with pedestrians and vehicles. Our approach runs on a Linux machine with an i7 16-core CPU, 16GB RAM, and an RTX 3080 GPU. It processes at around 60 Hz, with spatiotemporal graph generation at 300 Hz.

The qualitative results in normal traffic scenarios are shown in Figure 3. In these results, vehicle 1 receives query nodes from vehicle 2, and after CoID computation, vehicle 1 selects covisible objects (in green) to share back with vehicle 2. Non-covisible objects (in red) are outliers and should not be shared, as they negatively impact CoID performance and increase bandwidth usage. Our interactive CoID method effectively retrieves covisible objects by preserving the spatiotemporal relationships of street objects in a communication-efficient



(a) DGMC (b) Ours

Fig. 4. Comparison between DGMC and our approach on the F1 score given different number of iterations and bandwidth constraints.



(a) Temporal cues (b) CoID

Fig. 5. Analysis of our approach's characteristics in the CAD simulation, including the effect of the length of temporal sequence based on BIS and the improvements of CoID based on precision.

manner. The quantitative results in Table I show that our approach outperforms all the other methods in three scenarios in normal traffic. It is due to our approach's ability to integrate multi-vehicle spatial and temporal observations. Our approach achieves a significant improvement of $8\% - 56\%$ in covisible object retrieval and data-sharing efficiency based on the BIS metric, making it ideal for real-world applications with communication constraints. Additionally, our baseline method performs much better than other comparisons, highlighting the importance of temporal cue integration. The full approach surpasses the baseline due to our novel heterogeneous graph pooling and interactive CoID mechanism.

In Scenario 2, we conducted experiments to deeply compare our method with the existing DGMC method based on the F1 score, as depicted in Figure 4. The x and y axes represent the bandwidth and interaction number between connected vehicles. Our approach demonstrates superior performance compared to DGMC in both effectiveness and efficiency. From an effectiveness standpoint, both methods show gradual improvement with increasing bandwidth and interaction numbers. This is because of the increase of received query nodes from collaborators, which leads to the increase of recall. However, our approach consistently achieves a significantly higher F1 score compared to DGMC with the same number of interactions and bandwidth. From an efficiency standpoint, our approach attains its best performance by sharing/receiving around 10 nodes, as indicated by the yellow regions. In contrast, the DGMC method requires over 30 nodes to achieve a performance inferior to ours. As the vehicles start to share outliers in this case, it leads to a decrease in accuracy.

C. Results in Crowded Traffic Scenarios

We introduce a challenging Scenario 4 to thoroughly assess our approach. In contrast to Scenarios 1, 2, and 3, Scenario 4 features a higher density of entities, hosting over 20 street objects within a street intersection. Employing a sequence of frames to generate the spatiotemporal graph representation results in a graph with over 50 nodes. Scenario 4 introduces extensive interactions among street objects, significantly amplifying the complexity of the CoID task.

The quantitative results for crowded traffic scenarios in Table I show that our approach outperforms all metrics, thanks to the heterogeneous graph pooling technique, which efficiently selects covisible candidates from dense graphs. We achieve over a 25% improvement in covisible object retrieval and data-sharing efficiency (BIS metric), demonstrating the robustness of our approach in both normal and crowded scenes. Our baseline, Ours-NE, also surpasses other methods by integrating spatiotemporal cues for CoID. For the qualitative results, Figure 3 highlights objects observed by vehicle 1, showing that our method retrieves more covisible objects than DGMC, emphasizing the importance of spatiotemporal cue integration and interactive node sharing.

D. Discussion

Figure 5(a) depicts that the performance of our approach gradually decreases as the sequence length increases. When the sequence length is in the range of $[5, 6]$, the best performance is achieved. If the sequence length keeps increasing, the performance becomes stable with a small fluctuation, which also indicates the effectiveness of our approach which can capture the temporal cues for CoID.

Figure 5(b) presents the improvements of CoID accuracy by using our approach to retrieve covisible objects. We use our approach to retrieve covisible objects and then use the traditional graph-matching approach to identify the correspondences of objects. Without using our approach as a pre-process, the graph-matching performance is bad due to the existence of a large amount of outliers. Given the retrieved objects obtained by our approach, the CoID accuracy improves significantly, as our approach first shares the objects that are most likely to have correspondences, thus reducing a large number of outliers.

V. Conclusion

In this paper, we introduce a novel bandwidth-adaptive spatiotemporal CoID approach for collaborative perception. We formulate CoID as a spatiotemporal graph learning and matching problem, developing an interactive information-sharing paradigm that adapts to dynamic bandwidth constraints. Our method uses a heterogeneous graph attention network to integrate visual, spatial, and temporal features and employs graph pooling to select candidates for data sharing. Extensive experiments in both normal and crowded traffic simulations demonstrate that our approach achieves state-of-the-art CoID performance under varying bandwidth conditions.

REFERENCES

[1] L. Gallo and J. Härri, "A lte-direct broadcast mechanism for periodic vehicular safety communications," in *IEEE Vehicular Networking Conference*. IEEE, 2013, pp. 166–169.

[2] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *Intelligence Transportation System*, vol. 23, no. 3, pp. 1852–1864, 2020.

[3] D. Forsyth, "Object detection with discriminatively trained part-based models," *Computer*, vol. 47, no. 02, pp. 6–7, 2014.

[4] Z. Song, F. Wen, H. Zhang, and J. Li, "A cooperative perception system robust to localization errors," in *IEEE Intelligent Vehicles Symposium*, 2023.

[5] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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

[7] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *NIPS*, 2022.

[8] 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 *ACM/IEEE Symposium on Edge Computing*, 2019.

[9] J. Guo, D. Carrillo, S. Tang, Q. Chen, Q. Yang, S. Fu, X. Wang, N. Wang, and P. Palacharla, "Coff: Cooperative spatial feature fusion for 3-d object detection on autonomous vehicles," *Internet of Thing*, vol. 8, no. 14, pp. 11 078–11 087, 2021.

[10] 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 *European Conference on Computer Vision*, 2020.

[11] Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno, "Multi-robot collaborative perception with graph neural networks," *RAL*, vol. 7, no. 2, pp. 2289–2296, 2022.

[12] 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*, 2022.

[13] 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 *IEEE International Conference on Robotics and Automation*, 2022.

[14] R. Bi, J. Xiong, Y. Tian, Q. Li, and X. Liu, "Edge-cooperative privacy-preserving object detection over random point cloud shares for connected autonomous vehicles," *Intelligence Transportation System*, vol. 23, no. 12, pp. 24 979–24 990, 2022.

[15] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *NIPS*, 2021.

[16] 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*, 2022.

[17] Y. Yuan, H. Cheng, and M. Sester, "Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving," *RAL*, vol. 7, no. 2, pp. 3054–3061, 2022.

[18] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake lidar in 3d detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular slam," in *European Conference on Computer Vision*, 2014.

[20] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular slam system," *TRO*, vol. 31, no. 5, pp. 1147–1163, 2015.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2005.

[22] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[23] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *IEEE International Conference on 3D Digital Imaging and Modeling*, 2001.

[24] J. Zhang, Y. Liu, M. Wen, Y. Yue, H. Zhang, and D. Wang, "L2V2T2-Calib: Automatic and Unified Extrinsic Calibration Toolbox for Different 3D LiDAR, Visual Camera and Thermal Camera," in *IVS*, 2023.

[25] P. Gao, Z. Zhang, R. Guo, H. Lu, and H. Zhang, "Correspondence identification in collaborative robot perception through maximin hypergraph matching," in *IEEE International Conference on Robotics and Automation*, 2020.

[26] P. Gao, R. Guo, H. Lu, and H. Zhang, "Regularized graph matching for correspondence identification under uncertainty in collaborative perception," in *Robotics: Science and Systems*, 2021.

[27] K. Fathian, K. Khosoussi, Y. Tian, P. Lusk, and J. P. How, "Clear: A consistent lifting, embedding, and alignment rectification algorithm for multiview data association," *TRO*, vol. 36, no. 6, pp. 1686–1703, 2020.

[28] N. Hu, Q. Huang, B. Thibert, and L. J. Guibas, "Distributable Consistent Multi-Object Matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[29] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-Aligned Representation Learning for Person Re-identification," in *The AAAI Conference on Artificial Intelligence*, 2020.

[30] A. Khatun, S. Denman, S. Sridharan, and C. Fookes, "Semantic Consistency and Identity Mapping Multi-component Generative Adversarial Network for Person Re-identification," in *WACV*, 2020.

[31] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: Multi-object tracking and segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[32] R. Wang, J. Yan, and X. Yang, "Learning combinatorial embedding networks for deep graph matching," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[33] Z. Zhang and W. S. Lee, "Deep Graphical Feature Learning for the Feature Matching Problem," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[34] M. Fey, J. E. Lenssen, C. Morris, J. Masci, and N. M. Kriege, "Deep graph matching consensus," in *International Conference on Learning Representations*, 2019.

[35] P. Gao and H. Zhang, "Bayesian deep graph matching for correspondence identification in collaborative perception." in *Robotics: Science and Systems*, 2021.

[36] P. Gao, R. Guo, H. Lu, and H. Zhang, "Correspondence identification for collaborative multi-robot perception under uncertainty," *Autonomous Robots*, vol. 46, no. 1, pp. 5–20, 2022.

[37] E. Ranjan, S. Sanyal, and P. Talukdar, "Asap: Adaptive structure aware pooling for learning hierarchical graph representations," in *The AAAI Conference on Artificial Intelligence*, 2020.

[38] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning*, 2017.

[39] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "SUMO (simulation of urban mobility)-an open-source traffic simulation," in *The 4th Middle East Symposium on Simulation and Modelling*, 2002.

[40] M. Fey, J. E. Lenssen, F. Weichert, and H. Müller, "SplineCNN: Fast geometric deep learning with continuous b-spline kernels," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[41] M. Fey, J. E. Lenssen, C. Morris, J. Masci, and N. M. Kriege, "Deep graph matching consensus," in *International Conference on Learning Representations*, 2020.