IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

A Spatial Calibration Method for Robust Cooperative Perception

Zhiying Song, Tenghui Xie, Hailiang Zhang, Jiaxin Liu, Fuxi Wen, *Senior Member, IEEE*, Jun Li

***Abstract*—Cooperative perception is a promising technique for intelligent and connected vehicles through vehicle-to-everything (V2X) cooperation, provided that accurate pose information and relative pose transforms are available. Nevertheless, obtaining precise positioning information often entails high costs associated with navigation systems. Hence, it is required to calibrate relative pose information for multi-agent cooperative perception. This paper proposes a simple but effective object association approach named context-based matching (CBM), which identifies inter-agent object correspondences using intra-agent geometrical context. In detail, this method constructs contexts using the relative position of the detected bounding boxes, followed by local context matching and global consensus maximization. The optimal relative pose transform is estimated based on the matched correspondences, followed by cooperative perception fusion. Extensive experiments are conducted on both the simulated and real-world datasets. Even with larger inter-agent localization errors, high object association precision and decimeter-level relative pose calibration accuracy are achieved among the cooperating agents. Demo video, code, and more up-to-date information are available at https://github.com/zhyingS/CBM.**

***Index Terms*—Distributed Robot Systems, Object Detection, Pose Errors, Robustness.**

I. Introduction

COOPERATIVE perception has emerged as a prominent research topic for intelligent and connected vehicles in recent years [1]. This technique enhances the perception capability of the individual vehicles by leveraging complementary information from neighboring agents, *e.g.* , vehicles [2], drones [3] or infrastructure nodes [4].

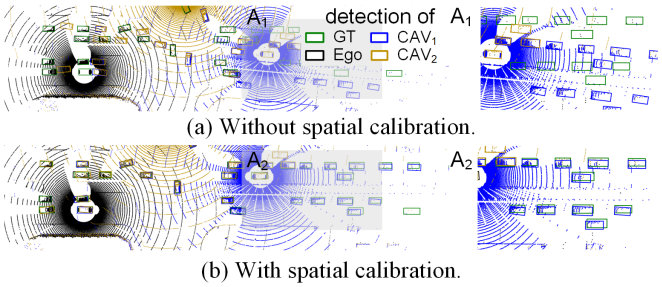
However, aligning perception results among these agents hinges on the availability of precise localization measurements, which can be challenging to obtain in complex traffic environments. Therefore, a spatial calibration module is needed to refine the spatial offset caused by localization errors [5]. As an illustration, Fig. 1 shows a cooperative perception scenario of three vehicles. The detection results are transformed into the Ego frame using transformation matrices acquired from localization systems. Errors in this transform result in significant misalignment of the detection results. In such a case, cooperative perception not only fails to improve Ego perception but also disrupt it.

Manuscript received: November, 9, 2023; Revised January, 26, 2024; Accepted February, 20, 2024.

This paper was recommended for publication by Editor M. Ani Hsieh upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Key R&D Program of China under Grant 2021YFB1600402 and 2020YFB1600303.

The authors are with the School of Vehicle and Mobility, Tsinghua University, Beijing, China. *Corresponding author*: Fuxi Wen, wenfuxi@tsinghua.edu.cn.

Digital Object Identifier (DOI): see top of this page.



(a) Without spatial calibration.

(b) With spatial calibration.

Fig. 1. Effect of spatial calibration on cooperative perception.

Previous research has suggested calibrating spatial errors through the alignment of raw data [6], [7], or the use of specific features [8], [9]. Nevertheless, in cooperative perception systems, the preference leans toward a lightweight method that requires minimal information transmission and involves the fewest additional feature extractors. In this paper, we propose to use only the object-level features, *i.e.* , the detection results in the form of bounding boxes, to achieve a robust calibration of spatial errors.

Current object-level calibration methods have faced difficulties owing to the following three challenges [10]–[12]. The first challenge is the scarcity of information. Object-level features are extracted by the object detection module, encompassing solely the position, orientation, and dimensions of the surrounding objects. In contrast to using raw data or deep features, this approach results in the loss of a considerable amount of semantic information. The second challenge is the presence of perception errors. The object-level features can be noisy due to the inherent limitations of the detection modules. For example, an object may be detected at an incorrect position with a significantly deviated heading angle. The perception noise further erodes the usability of object-level features. The third challenge lies in the non-co-visible objects. To achieve alignment of inter-agent features, it is crucial to identify the same object from different perspectives. However, a substantial portion of the objects are non-co-visible, meaning they are only visible to one of the cooperating agents. This results in a high proportion of outliers when performing inter-agent object association.

In this paper, we propose a novel inter-agent pose alignment module for object-level distributed cooperative perception systems. The core idea is to identify inter-agent objects by intra-agent geometrical context. The method consists of three steps. First, an intra-agent context matrix for each object is constructed by encoding their geometrical features. Then,

the unique correspondences between inter-agent objects are identified by seeking global consensus among the per-object context matrices. Finally, with these global correspondences, the relative transformation matrix is estimated and multi-view perception results are fused into an unified frame, resulting in a spatial error-calibrated cooperative perception output.

The proposed approach maximizes the utilization of features by embedding information from all objects into each object's local context matrix. Each context matrix, constructed from a local perspective, exhibits substantial distinctiveness, enhancing the method's resilience against outliers (non-co-visible objects). In addition, the perception errors are efficiently managed by seeking global consensus with redundancy among the context matrices of all the objects.

Our contributions are summarized as follows:

- We proposed a novel spatial calibration approach for distributed cooperative perception systems. Instead of assuming perfect spatial synchronization and ideal perception, we take localization and perception errors into account and design a system robust to them in complex traffic scenarios.
- We propose an effective inter-agent object association approach that is resilient to perception errors and outliers, achieved by taking into account the distinctive characteristics of objects within transportation scenarios.
- We achieve decimeter-level spatial calibration using only bounding boxes, rather than raw data or complex features. The proposed method can be naturally extended to V2X-based scenarios, with minimal communication overhead and cost-effective implementation.

II. Related Work

Cooperative perception. Recent research on multi-agent cooperative perception is mainly focused on improving efficiency, performance, robustness, and safety of the process [1], [13]. Significant progress in improving the detection performance of cooperative perception under ideal cases has been achieved in [2], [14]–[19]. For robustness, cooperative perception has been investigated for various issues, such as communication issues [20]–[22]. In terms of localization errors, V2VNet [2] was the first to demonstrate the sensitivity of cooperative perception to imperfect localization. Subsequently, many state-of-the-art cooperative perception models, such as those in [16], [19], [23], [24], have either demonstrated or emphasized this vulnerability. However, efficient solutions in this regard remain elusive.

Spatial calibration. Researchers tried to calibrate the localization errors using various features. Vadivelu *et al*. proposed a learning-based method to encode the sensor data to spatial feature maps and performed pose regression on them [8]. Yuan *et al*. selected bounding boxes, points of poles and points of big planar structures as features and developed a RANSAC-based inter-vehicle pose correction method [9]. Yang *et al*. proposed a feature descriptor for point cloud based on gridded Gaussian distribution with Wasserstein distance for global pose initialization [25]. TrajMatch calibrates inter-LiDAR pose at the roadside using trajectory and semantic features generated in the object detection/tracking phase [26]. However, these methods require the extraction of specific features for calibration, which represents an additional burden for cooperative perception systems. Several studies, such as [10], [27], have attempted to use only the results of perception systems and tackle inter-vehicle object association using lightweight point cloud registration algorithms. However, these methods often rely on a well-guessed initial pose and are unable to handle larger localization errors, thus limiting their effectiveness in complex scenarios.

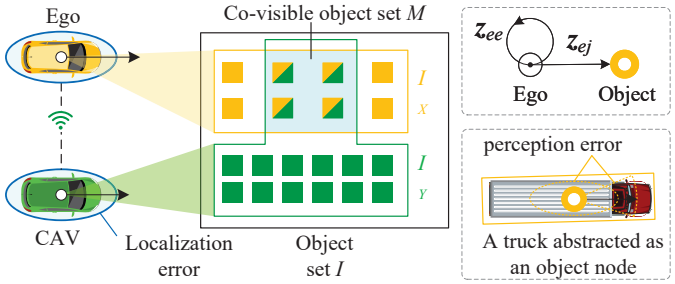


Fig. 2. Illustration of the object sets.

Object association. Iterative closest point (ICP) and its variants [28] are commonly used to associate dense object clusters, for example, in [29]. Nevertheless, the objects in cooperative perception are always distributed sparsely. Recently graph matching-related methods have gained popularity, they rely less on absolute positioning and instead use relative information between nodes for association. Gao *et al*. formulated the association problem in cooperative perception as a non-convex constrained graph optimization problem and developed a sampling-based algorithm to solve it [30]. However, the complexity and time-consuming nature of this approach hinder its applicability. Tedeschini *et al*. proposed to use a neural network to encode graph node and edge features for association [31]. A computationally efficient method named VIPS is proposed in [12] to solve the similar graph optimization problem that makes it available to infrastructure-assisted cooperative perception. However, they face challenges in relying on the design of similarity functions and constraint relaxation, handling perception errors and outliers, as well as tuning numerous hyperparameters.

The rest of the paper is organized as follows: In Section III, the spatial calibration approach is proposed. Section IV and V present the experimental evaluation of the real-world dataset SIND and simulated cooperative perception dataset OPV2V, respectively. Finally, Section VI summarizes the conclusions. The framework of the proposed method is shown in Fig. 3.

III. Method

A. Problem formulation

As illustrated in Fig. 2, consider a vehicular network consisting of two cooperative nodes $\mathcal{A} = \{e, c\}$ (one Ego agent and one cooperative agent, *e.g.* , connected vehicles or RSU) and some passive nodes denoted as objects $\mathcal{I}$ with cardinality N, which can be further segmented into two overlapped groups: objects $\mathcal{I}_\mathcal{X}$ with cardinality N_e that can

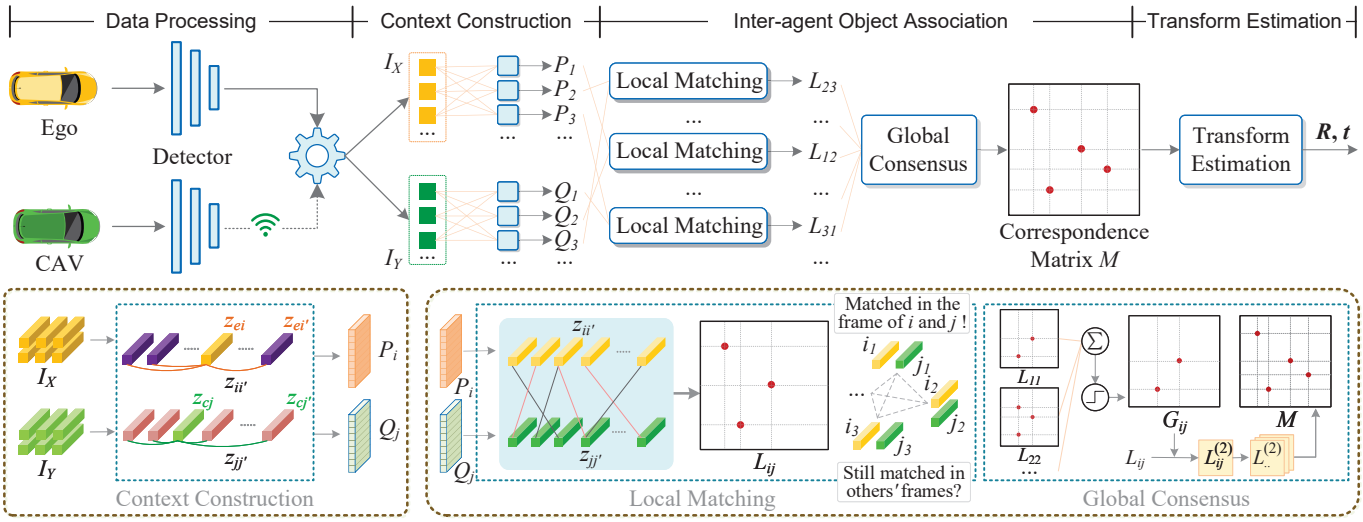


Fig. 3. Framework of the proposed method CBM. Object sets $\mathcal{I}_\mathcal{X}$ and $\mathcal{I}_\mathcal{Y}$ are detected by the onboard perception system of the Ego vehicle and CAV, respectively. Subsequently, the Ego establishes context based on $\mathcal{I}_\mathcal{X}$ and $\mathcal{I}_\mathcal{Y}$. Preliminary correspondences between objects in $\mathcal{I}_\mathcal{X}$ and $\mathcal{I}_\mathcal{Y}$ are identified via the local matching module, and subsequently refined through the global consensus module. Finally, the relative pose between the Ego and CAV is estimated.

be sensed by Ego agent and objects $\mathcal{I}_\mathcal{Y}$ with cardinality N_c that are accessible for sensing by the cooperative agent. In the presence of co-visible objects, the overlap between $\mathcal{I}_\mathcal{X}$ and $\mathcal{I}_\mathcal{Y}$ represents the co-visible object set $\mathcal{M}$.

The state of a node (either agent node or object node) defined in the world coordinate system is denoted by vector $\boldsymbol{s} = [\boldsymbol{p}^T, \boldsymbol{d}^T, \theta]^T \in \mathbb{R}^6$, which includes Bird's-eye-view position $\boldsymbol{p} \in \mathbb{R}^2$, orientation $\theta \in (0, 2\pi)$ and 3D dimension size (height, width and length) $\boldsymbol{d} \in \mathbb{R}^3$. We denote the measurement vector as

$$\boldsymbol{z}_{kh} = \left[\left(\boldsymbol{z}_{kh}^{\boldsymbol{p}}\right)^T, \left(\boldsymbol{z}_{kh}^{\boldsymbol{d}}\right)^T, \boldsymbol{z}_{kh}^{\theta} \right]^T \in \mathbb{R}^6, \forall\ k \in \mathcal{A}, h \in \mathcal{A} \cup \mathcal{I}$$

The notations $\boldsymbol{z}_{kh}$ for $k \neq h$ represent inter-node (relative) measurements, while $\boldsymbol{z}_{kk}$ represents the intra-node (absolute) measurements of agent $k \in \mathcal{A}$, as shown in Fig. 2.

Intra-node Measurements. The position and orientation states of agents are measured by their onboard localization and navigation systems, by

$$\boldsymbol{z}_k \triangleq \boldsymbol{z}_{kk} = \boldsymbol{g}^{(k)}(\boldsymbol{s}_k) + \boldsymbol{\omega}_k,\ \forall\ k \in \mathcal{A} \tag{1}$$

where $\boldsymbol{g}^{(k)}(\cdot)$ denotes a function of agent absolute localization states, and $\boldsymbol{\omega}_k$ represents the localization noise.

Inter-node Measurements. The states of the objects are measured by the perception systems of the agents, by

$$\boldsymbol{z}_{kh} = \boldsymbol{h}^{(k)}\left(\boldsymbol{s}_k, \boldsymbol{s}_h\right) + \boldsymbol{\omega}_{kh},\ \forall\ k \in \mathcal{A}, h \in \mathcal{I} \tag{2}$$

where $\boldsymbol{h}^{(k)}(\cdot)$ denotes the perception system, and $\boldsymbol{\omega}_{kh}$ represents the perception noise.

Given the above measurements, *i.e.* , the state measurement of both agents $\boldsymbol{z}_k$ and the detection measurement of objects $\boldsymbol{z}_{kh}$, the problem is to estimate the transformation matrix $\boldsymbol{T}_c^e$ between the coordinate of the Ego agent and the cooperative agent. We frame this problem as a measurement alignment issue, which is then subdivided into two components: inter-agent object association and transform estimation. The former is to find the matching set of co-visible objects. From a ground truth perspective, the object set $\mathcal{I}$ can be categorized into two groups: co-visible objects $\mathcal{M} = \mathcal{I}_\mathcal{X} \cap \mathcal{I}_\mathcal{Y}$ that can be jointly detected by both agents and non-co-visible objects $\overline{\mathcal{M}} = \mathcal{I} - \mathcal{M}$ that can only be sensed by one of the agents. However, these two sets are not directly observable from agents' local measurements. The inter-agent object association task is to estimate the co-visible object set $\hat{\mathcal{M}}$ from the local measurement sets, followed by the second task, *i.e.* , inter-agent transform estimation,

$$\hat{\boldsymbol{R}}_c^e, \hat{\boldsymbol{t}}_c^e = \arg\min || \left(\boldsymbol{R} \cdot \boldsymbol{z}_{ch}^{\boldsymbol{p}} + \boldsymbol{t}\right) - \boldsymbol{z}_{eh}^{\boldsymbol{p}} ||_2,\ h \in \hat{\mathcal{M}} \tag{3}$$

where $\hat{\boldsymbol{R}}_c^e \in SO(2)$ is a rotation matrix and $\hat{\boldsymbol{t}}_c^e \in \mathbb{R}^2$ denotes a translation vector.

B. Context-based inter-agent object association

Corresponding to the index set $\mathcal{I}_\mathcal{X}$ and $\mathcal{I}_\mathcal{Y}$ defined in section III-A, let the state of the object $i \in \mathcal{I}_\mathcal{X}$ be $\boldsymbol{s}_i = [\boldsymbol{p}_i^T, \boldsymbol{d}_i^T, \theta_i]^T \in \mathbb{R}^3$, where $\boldsymbol{p}_i \in \mathbb{R}^2$ and $\theta_i \in (0, 2\pi)$ is the 2D position and orientation in the Bird's-eye-view, respectively. $\boldsymbol{d}_i \in \mathbb{R}^3$ denotes 3D dimensions. Similarly, for object $j \in \mathcal{I}_\mathcal{Y}$, $\boldsymbol{s}_j = [\boldsymbol{p}_j^T, \boldsymbol{d}_j^T, \theta_j]^T$.

The goal of this subsection is to find the covisible object set $\hat{\mathcal{M}}$ that contains the same objects observed from different agents' view. A coarse-to-fine strategy is employed to approximate this matching correspondence. This involves initially identifying coarse matching sets that include possible results, followed by the removal of outliers and the attainment of a global consensus, ultimately yielding the final estimation. The details are shown in the following pages.

1) Intra-agent context construction: In real-world traffic environments, each traffic participant possesses distinct attributes, such as position, direction, and appearance. Con-

sequently, when viewed from the perspective of an individual vehicle, the surrounding environment is inherently unique, thereby constituting its context. In other words, context uniquely encodes the relationships between nearby objects from an object's local perspective. A simple case is shown in Fig. 4. Inspired by such an observation, we employ context-based comparisons to identify and locate identical objects across multiple views.

Similar to the preprocessing procedure in [10], we first standardize the measurements by converting them into the Ego frame using transform

$$\widetilde{\boldsymbol{T}}_c^e = f\left(\boldsymbol{z}_c, \boldsymbol{z}_e\right) \tag{4}$$

for $c, e \in \mathcal{A}$, where $\boldsymbol{z}_c$ and $\boldsymbol{z}_e$ are defined in (1), the transform function $f(\cdot)$ can be found in [10], representing the transform calculated from on-board localization systems. $\widetilde{\boldsymbol{T}}_c^e$ is the desired $\boldsymbol{T}_c^e$ if there is no localization noise in the measurement of intra-agent position and orientation. As a result,

$$\boldsymbol{z}_c^{(e)} = \widetilde{\boldsymbol{T}}_c^e(\boldsymbol{z}_c),\ \boldsymbol{z}_{cj}^{(e)} = \widetilde{\boldsymbol{T}}_c^e(\boldsymbol{z}_{cj}),\ \forall j \in \mathcal{I}_{\mathcal{Y}} \tag{5}$$

where the superscript $\cdot^{(e)}$ indicates that it's in the coordinate of the Ego agent. For brevity in the following text, we will omit the superscript, but it should be understood that all measurement values related to the cooperative agent have been transformed into the Ego coordinate system according to (5).

In the Ego frame, the directions of objects (both in $\mathcal{I}_{\mathcal{X}}$ and $\mathcal{I}_{\mathcal{Y}}$) are adopted by defining their heading towards the front in the Ego frame as the forward direction. Then the relative positional measurements between objects in the local frame are given by

$$\begin{aligned} \boldsymbol{z}_{ii'}^{\boldsymbol{p}} &= \boldsymbol{R}^T(\boldsymbol{z}_{ei}^{\theta})\left(\boldsymbol{z}_{ei'}^{\boldsymbol{p}} - \boldsymbol{z}_{ei}^{\boldsymbol{p}}\right), && \forall i, i' \in \mathcal{I}_{\mathcal{X}} \\ \boldsymbol{z}_{jj'}^{\boldsymbol{p}} &= \boldsymbol{R}^T(\boldsymbol{z}_{cj}^{\theta})\left(\boldsymbol{z}_{cj'}^{\boldsymbol{p}} - \boldsymbol{z}_{cj}^{\boldsymbol{p}}\right), && \forall j, j' \in \mathcal{I}_{\mathcal{Y}} \end{aligned} \tag{6}$$

where $\boldsymbol{R}(\theta) \in SO(2)$ is a rotation matrix of a rotation angle θ. Note that $\boldsymbol{z}_{ii'}^{\boldsymbol{p}}$ and $\boldsymbol{z}_{jj'}^{\boldsymbol{p}}$ contains perception errors in (2).

In real-world traffic scenarios, each traffic participant occupies a significant space considering their dimensions and the requirement of maintaining a safe distance between road users. Therefore, even with some measurement errors, their position vectors remain highly distinctive within their occupied space, *i.e.* , the discrimination of $\boldsymbol{z}_{ii'}^{\boldsymbol{p}}$ ($\boldsymbol{z}_{jj'}^{\boldsymbol{p}}$) as a feature vector can still be maintained, thereby we define it one of i's (j's) context vectors.

Incorporating all the objects in the vicinity, the context matrices are obtained,

$$\begin{aligned} \boldsymbol{P}_i &= [\boldsymbol{z}_{i1}^{\boldsymbol{p}}, \boldsymbol{z}_{i2}^{\boldsymbol{p}}, \ldots, \boldsymbol{z}_{ii'}^{\boldsymbol{p}}, \ldots] && \in \mathbb{R}^{2\times N_e} \\ \boldsymbol{Q}_j &= \left[\boldsymbol{z}_{j1}^{\boldsymbol{p}}, \boldsymbol{z}_{j2}^{\boldsymbol{p}}, \ldots, \boldsymbol{z}_{jj'}^{\boldsymbol{p}}, \ldots\right] && \in \mathbb{R}^{2\times N_c} \end{aligned} \tag{7}$$

As context captures the relationships among objects and their neighboring objects, it inherently includes robust spatial constraints between connected objects and remains invariant to rigid transformations. It is worth noting that the concept of context shares similarities with graph descriptors that encode intra-node and inter-node information. However, the context incorporates strong spatial constraints between connected objects, making it more suitable for real-world driving applications where spatial relationships play a crucial role.

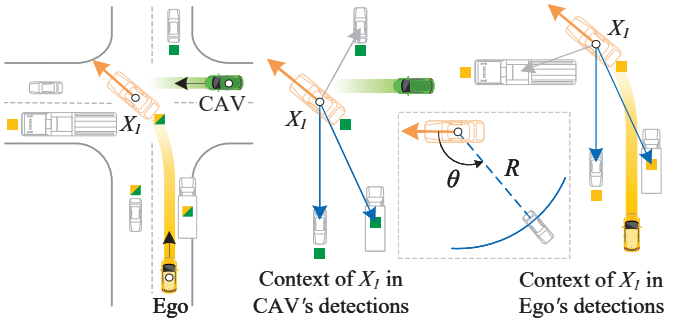


Fig. 4. Context of object X_1 in the detections of the Ego and CAV, respectively.

2) Context similarity-based coarse matching: Given $\boldsymbol{P}_i$ and $\boldsymbol{Q}_j$, the similarity between $\boldsymbol{z}_{ii'}^{\boldsymbol{p}}$ and $\boldsymbol{z}_{jj'}^{\boldsymbol{p}}$ (denoted as $\boldsymbol{z}_{i'}$ and $\boldsymbol{z}_{j'}$ below) is defined as

$$S_{\boldsymbol{z}} = \frac{\alpha}{\sigma_1}\left(\arccos\frac{\left|\boldsymbol{z}_{i'}^T\boldsymbol{z}_{j'}\right|}{\|\boldsymbol{z}_{i'}\|_2 \cdot \|\boldsymbol{z}_{j'}\|_2}\right) + \frac{\beta}{\sigma_2}\|\boldsymbol{z}_{i'} - \boldsymbol{z}_{j'}\|_1 \tag{8}$$

where $S_{\boldsymbol{z}} \in \mathbb{R}$, $||\cdot||_1$ denotes l_1 norm. The first term denotes the angular distance and the second characterizes the length difference between the local context of ith object in $\mathcal{I}_{\mathcal{X}}$ and jth object in $\mathcal{I}_{\mathcal{Y}}$. $\alpha > 0$ and $\beta > 0$ are the parameters to tune the weights of angular and length distance. σ_1 is set to tolerate angular perception errors caused by the positional error of the surrounding objects (i' and j') and the heading angle error of the center object (i and j). σ_2 is set to handle the vector length noise caused by the positional error of both the center and surrounding objects. The use of absolute value operation in the $\left|\boldsymbol{z}_{i'}^T\boldsymbol{z}_{j'}\right|$ term is intended to avoid ambiguity caused by heading direction since detecting the direction of a road user frequently results in opposite judgments.

For the sake of efficiency, we set $\alpha = 1$ and $\beta = 0$ first to get a preliminary similarity $S_{\boldsymbol{z}}^{(1)}$, then pick out those highly similar pairs to further compare Euclidean similarity $S_{\boldsymbol{z}}^{(2)}$. The whole procedure is defined as follows:

for $i \in \mathcal{I}_{\mathcal{X}}$ *and* $j \in \mathcal{I}_{\mathcal{Y}}$ **do**
  Initialize an void correspondence set $\mathcal{L}_{ij}$;
  for $i' \in \mathcal{I}_{\mathcal{X}}$ *and* $j' \in \mathcal{I}_{\mathcal{Y}}$ **do**
    if $S_{\boldsymbol{z}}^{(1)} = S_{\boldsymbol{z}}(\alpha = 1, \beta = 0) \leqslant 1$ *and* $S_{\boldsymbol{z}}^{(2)} = S_{\boldsymbol{z}}(\alpha = 0, \beta = 1) \leqslant 1$ **then**
      $\mathcal{L}_{ij} \leftarrow \mathcal{L}_{ij} \cup (i', j')$;
    end
  end
end

After these steps, a preliminary correspondence $\mathcal{L}_{ij}$ for each pair $i \in \mathcal{I}_{\mathcal{X}}$ and $j \in \mathcal{I}_{\mathcal{Y}}$ is obtained, resulting for a local correspondence matrix related to $\mathcal{L}_{ij}$ as

$$\boldsymbol{L}_{ij}(i', j') = \begin{cases} 1, & \text{if } (i', j') \in \mathcal{L}_{ij} \text{ and } \mathrm{card}(\mathcal{L}_{ij}) \geq 2. \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where $\boldsymbol{L}_{ij} \in \{0, 1\}^{N_e \times N_c}$, and operator $\mathrm{card}(\cdot)$ denotes counting the number of elements in the set.

By changing the soft thresholds σ_1 and σ_2, $\mathcal{L}_{ij}$ can encompass a large number of potential matches, aiming to include a significant portion of the ground truth correspondences. The solution of the object association problem can be obtained by filtering outliers from $\boldsymbol{L}_{ij}$, and we achieve this by maximizing the global consensus across all $i \in \mathcal{I}_{\mathcal{X}}$ and $j \in \mathcal{I}_{\mathcal{Y}}$.

3) Global consensus maximization: To filter out mismatched correspondences in $\boldsymbol{L}_{ij}$, a global filter matrix $\boldsymbol{G}_{ij} \in \{0,1\}^{N_e \times N_c}$ is developed for each object pair $i \in \mathcal{I}_{\mathcal{X}}$ and $j \in \mathcal{I}_{\mathcal{Y}}$. The basic idea is to assess the already matched pairs from a global perspective. We eliminate pairs that are accepted as matched ones in some objects' local frames but not embraced by all the objects,

$$\boldsymbol{G}_{ij}(i',j') = \begin{cases} 1, & \text{if } \sum_{(k',h') \in \mathcal{L}_{ij}} \boldsymbol{L}_{k'h'}(i',j') > 1. \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

This operation assesses the non-zero correspondences in $\boldsymbol{L}_{ij}$ from each other's perspective to maximize global consensus. Then the improved correspondence matrix becomes

$$\boldsymbol{L}_{ij}^{(1)} = \boldsymbol{G}_{ij} \circ \boldsymbol{L}_{ij}, \tag{11}$$

where $\boldsymbol{L}_{ij}^{(1)} \in \{0,1\}^{N_e \times N_c}$.

To further eliminate one-to-many matching correspondences, where one object in $\mathcal{I}_{\mathcal{X}}$ matches with several objects in $\mathcal{I}_{\mathcal{Y}}$, or vice versa, an extra rule is added, then

$$\boldsymbol{L}_{ij}^{(2)}(i',j') = \begin{cases} 0, \text{if } \begin{cases} \sum_{j'=1}^{N_c} \boldsymbol{L}_{ij}^{(1)}(i',j') > 1, \text{ or} \\ \sum_{i'=1}^{N_e} \boldsymbol{L}_{ij}^{(1)}(i',j') > 1. \end{cases} \\ \boldsymbol{L}_{i,j}^{(1)}(i',j'), \text{otherwise.} \end{cases} \tag{12}$$

Finally, the suboptimal matching correspondences can be obtained by

$$\boldsymbol{M} = \arg\max_{i,j} \left\| \boldsymbol{L}_{ij}^{(2)} \right\|_0 \tag{13}$$

where $||\cdot||_0$ is l_0 norm. The corresponding matched set is

$$\hat{\mathcal{M}} = \{(i,j) | i \in \mathcal{I}_{\mathcal{X}}, j \in \mathcal{I}_{\mathcal{Y}}, \boldsymbol{M}(i,j) \neq 0\}. \tag{14}$$

where i and j is the same object in real world.

C. Transform estimation and perception fusion

Given the set of matched object pairs $\forall\ (i,j) \in \hat{\mathcal{M}}$, if we denote the rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$, then

$$\boldsymbol{R}^*, \boldsymbol{t}^* = \arg\min_{\boldsymbol{R},\boldsymbol{t}} \sum_{i=1}^{|\hat{\mathcal{M}}|} \psi\left(\left\| \boldsymbol{z}_{ei}^{\boldsymbol{p}} - (\boldsymbol{R} \cdot \boldsymbol{z}_{ci}^{\boldsymbol{p}} + \boldsymbol{t}) \right\|_2 \right) \tag{15}$$

where $|\cdot|$ denotes the number of elements in the set, and $\boldsymbol{R} \in SO(2)$, $t \in \mathbb{R}^2$. We adopt the strategy in [11] to design $\psi(x)$ and solve $\boldsymbol{R}$ and $\boldsymbol{t}$, please check [11] for the details of the solution.

Finally, the estimated inter-agent transform matrix is

$$\hat{\boldsymbol{T}}_c^e = \begin{bmatrix} \boldsymbol{R}^* & \boldsymbol{t}^* \\ 0 & 1 \end{bmatrix} \cdot \widetilde{\boldsymbol{T}}_c^e \tag{16}$$

After imposing the calibration transform on the objects detected by the cooperative agent, the objects from multiple views are aligned under the Ego frame and fused using Non-maximum suppression (NMS) [15], [32], which is typically integrated as the final step of the object detection algorithm. Please refer to [32] for details.

IV. Validation on Real-World Dataset

A. Experiments setting

Dataset. Due to the nascent stage of cooperative perception technology, most datasets are focused on evaluating object detection performance, and very few datasets are available for evaluating spatial robustness and object association performance. We opt to use SIND [33], which is a real-world drone dataset captured from a signalized intersection from a stationary aerial perspective for about 420 minutes. The dataset includes more than 13,000 traffic participants in various types like cars, pedestrians, and motorcycles.

Metrics and Benchmarks. Given the estimated association set $\hat{\mathcal{M}}$ and the ground truth matching set $\mathcal{M}$, we evaluate the average precision. Three benchmarks are considered, including Iterative Closest Point (ICP) [28], Robust Iterative Closest Point (RICP) [11], and VIPS [12]. ICP is a fundamental technique for point association. As a classical method, many variants occurred in recent years, among which RICP is the latest achievement. VIPS is the state-of-the-art method for inter-vehicle object association using graph matching techniques. Compared with other graph matching-based algorithms, faster processing speed and higher accuracy are achieved for VIPS.

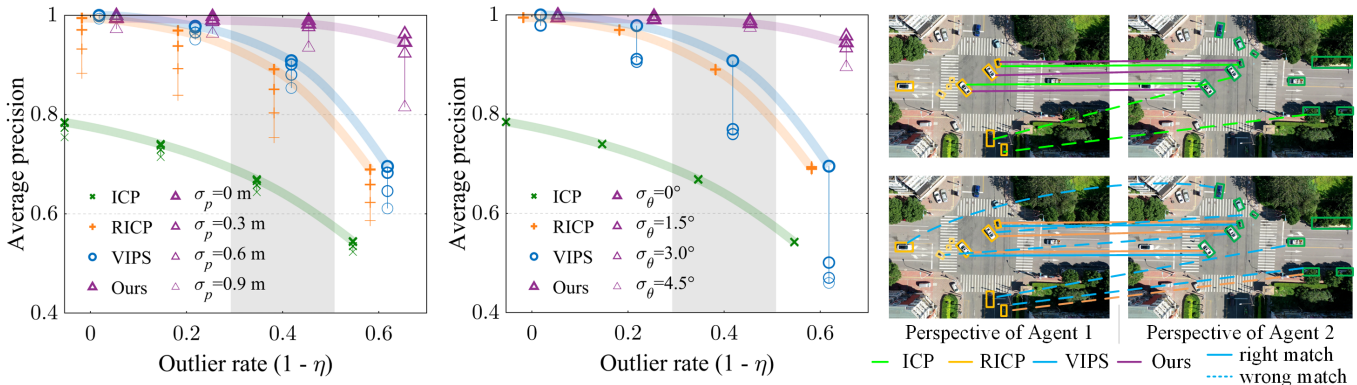
Co-visible objects. In real-world traffic scenarios, non-covisible objects exist due to a limited field of view and occlusions. These objects are outliers that have a serious impact on matching tasks. Given the object index set $\mathcal{I}$ at a single frame, we randomly sampled the co-visible object set $\mathcal{M}$ to simulate cooperative perception, such that $\text{card}(\mathcal{M}) = \eta \cdot \text{card}(\mathcal{I})$, $\mathcal{M} \subseteq \mathcal{I}$, where η is the rate of co-visible objects. The remaining objects are evenly assigned to the two cooperative agents, then we have $\mathcal{I}_{\mathcal{X}} \cup \mathcal{I}_{\mathcal{Y}} = \mathcal{I}$, $\mathcal{I}_{\mathcal{X}} \cap \mathcal{I}_{\mathcal{Y}} = \mathcal{M}$, where $\mathcal{I}_{\mathcal{X}}$ and $\mathcal{I}_{\mathcal{Y}}$ denote the perceived set by the two agents.

Perception errors. To investigate the impact of perception errors in (2), different levels of position and orientation angle errors are added to the objects in the dataset. They are set to be Gaussian distributed as $\mathcal{N}(0, \sigma_p)$ and $\mathcal{N}(0, \sigma_\theta)$, respectively. For object detection algorithms, determining the orientation of an object is a difficult task and prone to errors. To simulate this, we added a direction noise to the orientation with a 50% probability to make it face the opposite orientation.

Localization errors. Since the initial relative pose transformation reflects the magnitude of the pose error of the cooperating vehicles, it is set as a fixed value. In practice, the objects in $\mathcal{I}_{\mathcal{Y}}$ are translated by 3 m in the x and y directions and rotated by $5°$ as a whole, *i.e.* , the agents' relative position offset entirely based on accurate poses of the two vehicles.

B. Average precision of inter-agent object association

We test the performance of benchmarks on inter-agent object association under different levels of outlier rate and



(a) AP under different position errors. (b) AP under different orientation errors. (c) An example.

Fig. 5. Quantitative results and qualitative demonstration on SIND.

perception errors (including position errors and orientation errors), the results are shown in Fig. 5.

As shown in the result, η has a more significant impact than standard deviations σ_p and σ_θ. RICP exhibits a higher overall AP level than ICP, but they are both highly sensitive to η, this might be due to their use of iterative searching for the closest point in the correspondence identification step that converged to a local optimum. VIPS outperforms them in terms of AP, and it shows good robustness to changes in η. The proposed method outperforms the previous three methods in both overall precision and robustness to changes in η.

When considering the perception errors, we observed that the proposed algorithm is robust to position and orientation errors, with only an overall downward shift in the AP curve at $\sigma_p = 0.9$ m. For different levels of errors, the AP curve remains highly consistent with the zero error scenarios. RICP exhibits poor robustness to position errors, VIPS demonstrates good robustness against position errors but is unable to deal with large orientation errors. This is because VIPS uses the sine difference of the heading angles of two nodes to encode edge-to-edge similarity. This makes it fragile to errors contained in the heading angle of the objects.

V. Evaluation on Cooperative Perception Dataset

A. Experiment setting

Dataset. OPV2V [15] is a large-scale dataset that contains 73 scenes for V2V-based cooperative perception, including 2170 frames for *test* subset, and 549 frames for *test culver city* (*tcc*). The latter is developed to narrow down the gap between the simulated and real-world traffic scenarios, which can be used to test the adaptability and portability of the proposed algorithm. The reasons of using OPV2V are *a*) incorporation of cooperative vehicles and their locally perceived information, and *b*) provision of a wide range of scenarios, including highly complex traffic scenes with numerous traffic participants. The first row of Fig. 6 illustrates the distributions of co-visible object rate and absolute object counts across two test sets.

Object detection and perception errors. Object detection module provides inter-node measurements in (2). For fairness, We trained an object detection network PointPillars [34] using the *train* subset provided by OPV2V and kept it the same

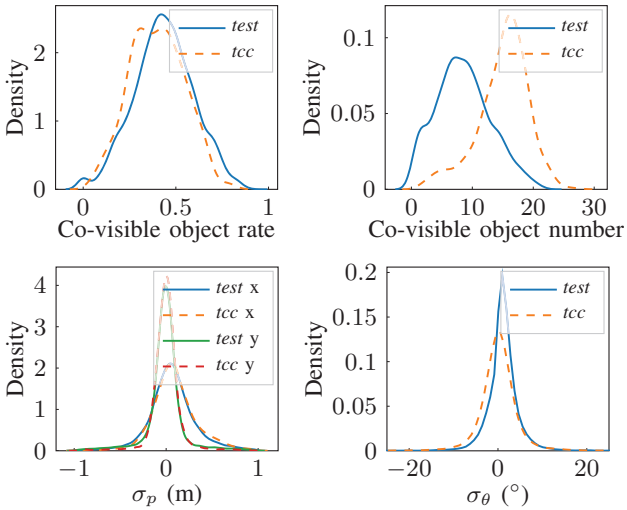


Fig. 6. Statistics of OPV2V dataset.

for all benchmarks. The performance of PointPillars on the dataset is evaluated in the second row of Fig.6. The third subfigure of Fig. 6 depicts the distribution of lateral (y) and longitudinal (x) position errors in the bounding boxes detected by PointPillars. It shows that the position errors in both directions approximately follow a Gaussian distribution, with greater errors observed in the longitudinal direction, and similar error distributions are observed across both datasets. This result supports the setting of σ_p in SIND. The fourth subfigure of Fig. 6 shows the distribution of angular errors in the network's perceived results. Specifically, we calculated the degree of deviation between the perceived bounding boxes and ground truth in the heading direction. It shows that the angular errors were distributed mostly between $-20°$ and $20°$, posing a significant challenge to association algorithms.

Localization errors. Without loss of generality, two scenarios are considered for demonstration and comparison: the first one assumes that the participating agents have no position and orientation errors, while the second one assumes that the position and orientation errors both follow zero-mean Gaussian distribution with standard deviation $\sigma_p^L = 3$ m and $\sigma_\theta^L = 5°$.

B. Evaluation of inter-agent association performance

Precision and recall. The precision and recall results for the matching task on OPV2V are shown in Table. I. Compared with the benchmarks, the proposed method achieves higher precision and recall and shows good robustness to the pose errors. Note that VIPS performs significantly worse on OPV2V compared to SIND, primarily due to the complex and challenging natures of the scenarios in OPV2V, such as a larger amount of objects and perception errors.

Distance between correspondence pairs. Matching precision and recall may not be a perfect indicator of the perception performance, for example, when the two objects are close to each other, their incorrect pose estimation would not deviate significantly from the ground truth, it would not have a major impact on the perception results. Therefore, another metric that can assess the impact of matching performance on perception is required. We defined a metric AD (average distance) $d = 1/N \sum_{(i,j)\in\hat{\mathcal{M}}} \mathrm{d}(\boldsymbol{s}_i - \boldsymbol{s}_j)$ that measures the average distance between the matched object pairs, where $N = \mathrm{card}(\hat{\mathcal{M}})$, and operator $\mathrm{d}(\cdot)$ denotes calculating the Euclidean distance. Table. I shows the performance of the methods of d on two datasets, *test* and *test culver city*. It's notable that the d values of the proposed method are quite small, which means even for incorrectly associated object correspondences in the matching results, the distances between them are not too far away to cause fatal impacts on the estimation of the pose transformation in the back end.

Impact of outliers. We present Fig. 7 depicting the distribution of average matching precision against the rate of non-co-visible objects on the *test culver city* and *test* datasets. The results consistently align with those obtained from the SIND dataset, as illustrated in Fig. 5. This reaffirms that the proposed method exhibits remarkable resilience to outliers.

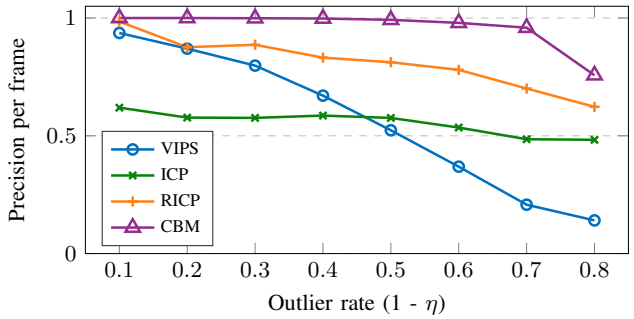


Fig. 7. Impact of outliers.

Impact of global consensus module. Ablation study is implemented by substituting $\mathbf{L}_{ij}^{(2)}$ in (13) with $\mathbf{L}_{ij}$, the result is shown in the last row of Table I. It's observed that the absence of the global consensus module has a minor impact on precision and recall. However, it has a significant influence on the matching distance, showing an increase of approximately 35% and 50% on the two datasets, respectively. This indicates that global consensus is effective to distinguish those mismatches with ambiguity.

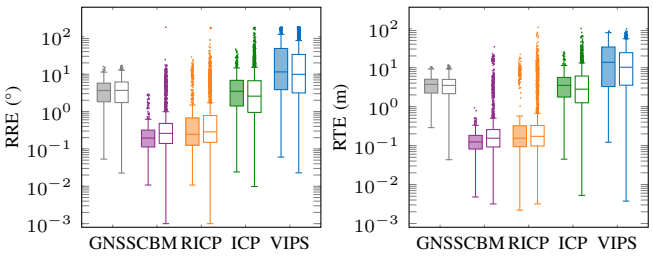


Fig. 8. Distribution of RRE and RTE on OPV2V.

TABLE I: Matching performance on OPV2V with $\sigma_p^L = 3$ m, $\sigma_\theta^L = 5°$. Pre.: precision, Rec.: recall, Dis.: distance. Ablation: without global consensus module.

Dataset	Metric	Method				
		VIPS	ICP	RICP	CBM	Ablation
tcc	Pre. (%)	47.8	46.0	73.6	**99.5**	-3.1
	Rec. (%)	58.5	38.4	72.1	**93.0**	-0.3
	Dis. (m)	28.27	7.84	3.99	**0.32**	+0.11
test	Pre. (%)	47.2	56.1	78.1	**90.6**	-5.1
	Rec. (%)	57.5	52.2	**80.8**	78.0	-0.8
	Dis. (m)	22.32	5.53	2.58	**0.50**	+0.25

TABLE II: mean Average precision at IoU=0.7 on OPV2V under different σ_p^L.

Dataset	Method	σ_p^L (m)					
		0	0.6	1.2	1.8	2.4	3.0
test	Single	60.0	60.0	60.0	60.0	60.0	60.0
	GNSS	**80.5**	24.5	21.4	24.2	26.0	26.5
	ICP	44.3	37.3	32.6	28.1	26.5	26.8
	VIPS	25.9	24.2	23.5	22.8	22.8	22.9
	RICP	71.9	**71.6**	**70.9**	**68.9**	67.4	64.3
	CBM	70.3	68.9	68.8	68.8	**68.9**	**68.9**
tcc	Single	47.0	47.0	47.0	47.0	47.0	47.0
	GNSS	**68.4**	28.8	25.8	25.3	25.5	25.9
	ICP	35.8	32.2	29.3	28.3	26.9	26.8
	VIPS	25.7	25.8	25.6	25.7	25.8	25.8
	RICP	62.3	**62.3**	**62.1**	59.7	57.8	51.7
	CBM	62.1	61.9	61.9	**61.9**	**61.9**	**61.9**

C. Evaluation of transform estimation and perception

The evaluation metrics for transform estimation are *a*) relative rotation error RRE $= \arccos\left(0.5 \cdot \mathrm{Tr}\left(\boldsymbol{R}^T \cdot \hat{\boldsymbol{R}}\right) - 0.5\right)$, and *b*) relative translation error RTE $= \|\boldsymbol{t} - \hat{\boldsymbol{t}}\|_2$. Heading errors are added to the cooperative agents for RRE estimation, and only position errors are introduced for the evaluation of RTE. The benchmark performances are compared at a fixed localization error level $\sigma_p^L = 3$ m or $\sigma_\theta^L = 5°$. The results are shown in Fig. 8, where the GNSS benchmark corresponds to the results without calibration. The proposed method has significantly reduced the median RRE to the level of $0.1°$ and the median RTE to the level of 0.1 m, achieving an order of magnitude improvement over the GNSS solution and outperforming other benchmarks.

Furthermore, perception robustness of the methods against pose errors is evaluated. The evaluation metric is mean Average Perception (mAP), computed by comparing the Intersec-

tion over Union (IoU) of fused bounding boxes and the ground truth boxes. An IoU threshold of 0.7 is chosen. Note that this metric is different to average precision used in Sec. V-B. The results are presented in Table II. Due to page limitations, only results for different σ_p^L are shown. However, similar trends can be observed in the results for different σ_θ^L. The proposed method not only maintains a high level of performance in mAP, also it is insensitive to agent pose errors under different levels, aligning with the results in Fig. 8.

VI. Conclusions

We propose a novel object-level spatial calibration approach for connected and automated driving to address the challenges of obtaining accurate relative transformation with dynamic and random position and pose errors. The proposed method enables robust inter-agent object association and relative pose estimation, leading to improved object-level cooperative perception. Its performance is demonstrated by extensive evaluations on the real-world dataset SIND and the cooperative perception dataset OPV2V.

References

[1] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "V2x cooperative perception for autonomous driving: Recent advances and challenges," *arXiv preprint arXiv:2310.03525*, 2023.

[2] 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 (ECCV)*. Springer-Verlag, 2020, p. 605–621.

[3] 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.

[4] M. Khamooshi, "Cooperative vehicle perception and localization using infrastructure-based sensor nodes," 2023.

[5] A. Caillot, S. Ouerghi, P. Vasseur, R. Boutteau, and Y. Dupuis, "Survey on cooperative perception in an automotive context," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 204–14 223, 2022.

[6] S. Kim, H. Kim, W. Yoo, and K. Huh, "Sensor fusion algorithm design in detecting vehicles using laser scanner and stereo vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1072–1084, 2015.

[7] S. Fang, H. Li, and M. Yang, "LiDAR SLAM based multivehicle cooperative localization using iterated split CIF," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21 137–21 147, 2022.

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

[9] Y. Yuan and M. Sester, "Leveraging dynamic objects for relative localization correction in a connected autonomous vehicle network," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, pp. 101–109, 2022.

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

[11] J. Zhang, Y. Yao, and B. Deng, "Fast and robust iterative closest point," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3450–3466, 2022.

[12] S. Shi, J. Cui, Z. Jiang, Z. Yan, G. Xing, J. Niu, and Z. Ouyang, "VIPS: real-time perception fusion for infrastructure-assisted autonomous driving," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking (MobiCom)*, 2022, pp. 133–146.

[13] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets and challenges," *arXiv preprint arXiv:2301.06262*, 2023.

[14] 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 (SEC)*, 2019, p. 88–100.

[15] 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 (ICRA)*, 2022.

[16] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Advances in Neural Information Processing Systems (NIPS)*, 2022.

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

[18] L. Yang, K. Yu, T. Tang, K. Yuan, L. Wang, X. Zhang, and P. Chen, "BEVHeight: A robust framework for vision-based roadside 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] B. Zhao, W. ZHANG, and Z. Zou, "Bm2cp: Efficient collaborative perception with lidar-camera modalities," in *7th Annual Conference on Robot Learning*, 2023.

[20] J. Tu, T. Wang, J. Wang, S. Manivasagam, M. Ren, and R. Urtasun, "Adversarial attacks on multi-agent communication," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7768–7777.

[21] J. Wang, Z. Wang, B. Yu, J. Tang, S. L. Song, C. Liu, and Y. Hu, "Data fusion in infrastructure-augmented autonomous driving system: Why? where? and how?" *IEEE Internet of Things Journal*, 2023.

[22] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles*, 2023.

[23] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[24] K. Yang, D. Yang, J. Zhang, M. Li, Y. Liu, J. Liu, H. Wang, P. Sun, and L. Song, "Spatio-temporal domain awareness for multi-agent collaborative perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 383–23 392.

[25] C. Yang, Z. Zhou, H. Zhuang, C. Wang, and M. Yang, "Global pose initialization based on gridded Gaussian distribution with Wasserstein distance," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2023.

[26] H. Ren, S. Zhang, S. Li, Y. Li, X. Li, J. Ji, Y. Zhang, and Y. Zhang, "Trajmatch: Towards automatic spatio-temporal calibration for roadside lidars through trajectory matching," *arXiv preprint arXiv:2302.02157*, 2023.

[27] A. Rauch, S. Maier, F. Klanner, and K. Dietmayer, "Inter-vehicle object association for cooperative perception systems," in *International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2013, pp. 893–898.

[28] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[29] J. Dong, Q. Chen, D. Qu, H. Lu, A. Ganlath, Q. Yang, S. Chen, and S. Labi, "Lidar-based cooperative relative localization," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.

[30] P. Gao, R. Guo, H. Lu, and H. Z. Zhang, "Regularized graph matching for correspondence identification under uncertainty in collaborative perception," in *Robotics science and systems (RSS)*, 2021.

[31] B. C. Tedeschini, M. Brambilla, L. Barbieri, and M. Nicoli, "Addressing data association by message passing over graph neural networks," in *International Conference on Information Fusion (FUSION)*, 2022, pp. 01–07.

[32] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS — improving object detection with one line of code," *IEEE International Conference on Computer Vision (ICCV)*, pp. 5562–5570, 2017.

[33] Y. Xu, W. Shao, J. Li, K. Yang, W. Wang, H. Huang, C. Lv, and H. Wang, "SIND: A drone dataset at signalized intersection in china," in *IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 2471–2478.

[34] 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 (CVPR)*, 2019.