V2X-Reg++: A Real-time Global Registration Method for Multi-end Sensing System in Urban Intersections

Xinyu Zhang[1,2,3,4,*], Qianxin Qu[1,*], Yijin Xiong[1,*,†], Chen Xia[1], Ziqiang Song[1], Qian Peng[1], Kang Liu[1], Jun Li[1], Keqiang Li[1]

***Abstract*—Urban intersections, dense with pedestrian and vehicular traffic and compounded by positioning signal obstructions, are among the most challenging areas in urban traffic systems. Traditional single-vehicle intelligence systems often perform poorly in such environments due to a lack of global scene observations and the inherent uncertainty in predicting other agents' intentions. Vehicle-to-Everything (V2X) technology, through real-time communication between vehicles (V2V) and vehicles to infrastructure (V2I), offers a robust solution. However, practical applications still face numerous challenges. Spatial registration among vehicle and infrastructure endpoints with different configurations in multi-end sensing systems is crucial for ensuring the accuracy of perception system data. Most existing multi-end spatial registration methods rely on initial extrinsic values provided by positioning systems, but the instability of GNSS signals due to high buildings in urban canyons poses severe challenges to these methods. To address this issue, this paper proposes a novel multi-end spatial registration method that does not require positioning priors to determine initial external parameters and meets real-time requirements. Our method introduces an innovative multi-end perception object association technique that leverages a new *Overall Distance* (*oDist*) metric to measure the spatial association between perception objects, subsequently using this metric as the foundation for an optimal transport formulation. By this means, we can extract co-observed targets from object association results for further external parameter computation and optimization. Extensive comparative and ablation experiments conducted on the simulated dataset V2X-Sim and the real dataset DAIR-V2X confirm the effectiveness and efficiency of our method. The code for this method can be accessed at: https://github.com/MassimoQu/v2i-calib.**

I. INTRODUCTION

Urban intersections, as key nodes of urban transportation networks, handle significant volumes of pedestrian and vehicle traffic daily. Particularly in city centers and commercial districts, the road junctions at these sites exhibit highly dynamic traffic flows, not only increasing the risk of traffic accidents [1] but also posing substantial challenges to traditional traffic management systems.

This research has been supported by the National Natural Science Foundation of China (Project No. 52221005; 62273198), and the Beijing Natural Science Foundation Program under Grant No.L241017, and the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG2-GC-2023-007).

* These authors contributed equally to this work as co-first authors.

† Corresponding author: Yijin Xiong(yj-xiong@mail.tsinghua.edu.cn).

[1]The State Key Laboratory of Automotive Safety and Energy, and the School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China.

[2]Tsinghua Automotive Research Institute (Suzhou), Tsinghua University, Suzhou, China.

[3]Electrical and Computer Engineering, National University of Singapore, Singapore, 117583.

[4]Anhui Mengshi Aviation Technology Co., Ltd., Hefei, China.

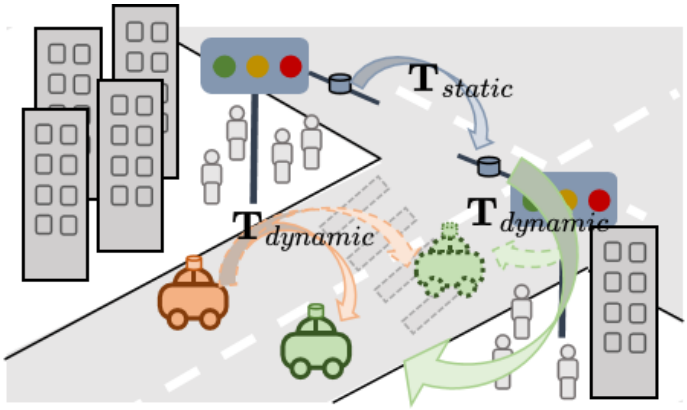


Fig. 1: Schematic of extrinsic parameters challenges at urban intersections.

In complex environments, conventional intelligent transportation systems often struggle to adapt due to a lack of a holistic view of traffic conditions and an inability to proactively model predict traffic flow evolution [2]. To address these challenges, Vehicle-to-Everything (V2X) technology [3] integrates real-time data exchange between vehicles (V2V) [4] and infrastructure (V2I) [5], enabling efficient management of traffic scenarios, such as congestion and accidents [1], and significantly improving urban traffic safety and efficiency.

However, this technology also introduces new challenges, particularly in the areas of multi-sensor data fusion and registration. In V2X systems, the multi-end sensing system, typically consisting of multiple LiDAR devices [6]–[9], requires precise spatial registration to ensure data consistency and accuracy. Due to the high dynamism of urban intersection scenarios, traditional static single-time spatial calibration methods [10] no longer meet practical needs. Moreover, existing multi-end spatial registration techniques [8], [9], [11]–[15] generally rely on positioning systems (e.g., GNSS) to provide high-precision initial extrinsic values. However, in practice, it is challenging for positioning systems to consistently meet the requirement, limiting the applicability of such methods in real-world scenarios.

In urban environments, the urban canyon effect often causes fluctuations in GNSS signals due to building obstructions and signal reflections. This issue is particularly pronounced at urban intersections, where low vehicle speeds make it difficult to distinguish effective signals from interference [16], further exacerbating positioning instability. The US Department of Transportation's V2X project [17] identified insufficient positioning accuracy at intersections as a major challenge.

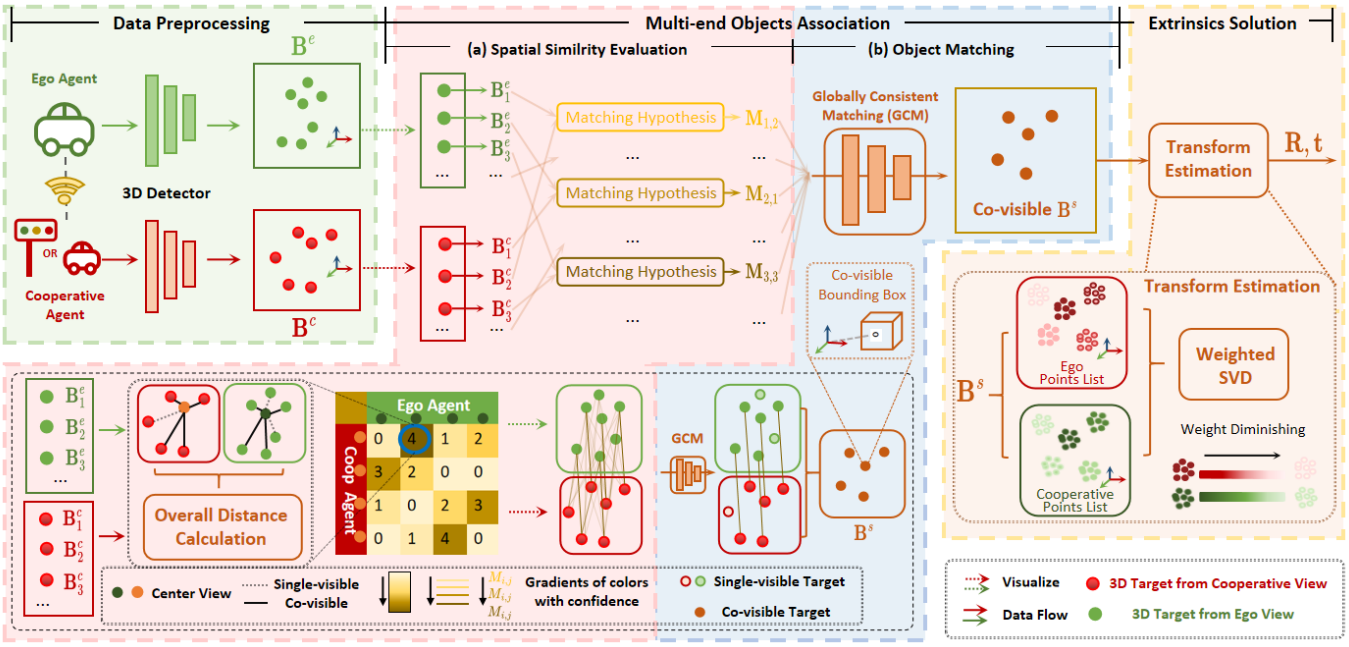


Fig. 2: The proposed method first generates 3D detection boxes on each endpoint, then identifies the common objects across endpoints. This multi-end object association is achieved through two steps: spatial similarity evaluation and object matching. The core process involves filtering the association of multi-end 3D detection boxes based on the affinity matrix (evaluated through *oDist* Calculation, with visualizations of both the matrix and the association degree). After identifying the common objects, feature point clouds are extracted from the detection boxes, and the extrinsic parameters are further computed using weighted point cloud registration.

Related experiments [18] also revealed that "missing relative positioning data" is notably more frequent in intersection scenarios, highlighting the difficulty of ensuring reliable positioning accuracy even with advanced systems. In addition, malicious attacks targeting V2X positioning systems [19]–[21] pose another significant threat to their stability at urban intersections. These challenges create a fundamental conflict between the positioning demands and inherent instability of urban intersections, making this a critical bottleneck for the widespread adoption of V2X technologies.

To address this issue, this paper introduces V2X-Reg++, an real-time global spatial registration method for multi-end sensing systems that operates without external positioning support. By matching detected objects from vehicle and roadside sensors, our method provides the real-time data alignment essential for reliable V2X applications.

The core of our method is a strategic shift from direct cross-source sensor data processing to leveraging the spatial topology of 3D detected objects. This perception-guided approach dramatically reduces computational complexity and enhances robustness to noise and partial overlaps typical in V2X environments. The method consists of two main components: a multi-end target association algorithm and an external parameter solution based on shared target . In the multi-end target association component, we introduce a novel *Overall Distance* (*oDist*) metric. This metric is informed by initial SVD-based transformation hypotheses for candidate object pairs and quantifies scene-level spatial consistency. This rich *oDist* measure then serves as the basis for an Optimal Transport (OT) formulation to achieve robust object matching. Subsequently, the feature point clouds derived from these confidently matched objects are used in a weighted SVD algorithm, where weights reflect matching confidence, to derive the final extrinsic parameters.

While classical spatial registration techniques [22] often rely on iterative closest point algorithms with SVD for refinement or apply OT to point-level correspondences, these approaches struggle with the high outlier ratios, lack of initial alignment, and scene-level complexity inherent in V2X scenarios. V2X-Reg++ differentiates itself by operating at the semantic level of detected objects, which is inherently suited to tackling the aforementioned V2X challenges. Methodologically, its key innovation is the synergistic framework of SVD and OT: SVD is not merely a final solver but an initial engine for generating object-pair transformation hypotheses that underpin our *oDist* metric. This *oDist* then empowers a more robust and meaningful OT stage for global object association, a distinct departure from traditional point-based applications of these techniques.

The main advantages of V2X-Reg++ is its independence from initial external parameters without compromising its real-time performance. Furthermore, this method leverages traffic participant information commonly present in traffic scenes, enhancing its versatility. By processing information using only perception data from target detection, it elegantly addresses the challenge of high outlier ratios in cross-source point clouds arising from different sensor configurations. Compared to other methods requiring complex data processing [13], V2X-

Reg++ has lower computational complexity and data transmission costs, making it more suitable for practical applications.

The innovations of this paper are summarized as follows:

1) An initial-value-free real-time spatial registration method for vehicle-road multi-end sensing system is proposed, particularly suited for environments like urban canyons where positioning fails;
2) A new multi-end target association method is proposed, which robustly establishes spatial associations across the scene even without positioning priors, and its core metric, *Overall Distance* (*oDist*), serves as a real-time indicator of external parameters quality among all participants;
3) The effectiveness of the method is validated on both simulated and real datasets, achieving real-time registration of external parameters.

In the preliminary conference paper [23], we proposed a framework for initial-value-free multi-end LiDARs registration. This article builds upon our preliminary work by introducing a more robust indicator of scene consistency, a novel weighting mechanism, an application example, and more comprehensive experimental validation. Notably, we have refined our core *oDist* metric (see Section IV-B2), superseding our previous *oIoU* metric, yielding greater computational efficiency (see time comparisons in Table II and Table III) and providing a more robust indicator of scene consistency (referencing the experiment in Fig. 8), further substantiated by new real-world application examples (Section VI). Additionally, V2X-Reg++ incorporates a novel weighted mechanism into the final extrinsic parameter estimation stage (see Section IV-C), which utilizes association confidences to significantly boost registration accuracy. The system's capabilities and resilience are also more comprehensively demonstrated through expanded experimentation, including evaluations on a new simulated V2X-Sim dataset (see Table II) and in-depth comparative analysis (see Table III). Collectively, these enhancements deliver a method with demonstrably higher accuracy and efficiency for real-time global registration in complex V2X environments.

II. RELATED WORK

This paper focuses on the multi-end spatial registration method in urban intersection scenarios. The core objective of this method is to solve the extrinsic parameters between different sensors, i.e., to determine their relative pose relationships, thereby achieving the spatial registration of multi-end sensing system.

Some existing registration methods can be applied to the scenario described in this paper. We categorize these methods into three groups: Target-based, Targetless, and Learning-based methods.

A. Target-Based Methods

Target-based methods typically involve placing calibration targets or boards with known dimensions and geometric features in the scene [24], and achieving high-precision extrinsic parameter estimation by detecting these corresponding features. The advantage of such methods lies in the straightforward extraction and matching of local features for registration, as well as controllable registration accuracy. However, in large-scale dynamic settings such as urban intersections [8], calibration targets may be occluded or cannot be left in place for an extended period. This inherent limitation underscores their offline nature, making them unsuitable for the demands of online registration and impractical for real-world deployment.

B. Targetless Methods

Targetless approaches do not rely on additional calibration targets and can be further divided into motion-based and scene-based methods:

Motion-based methods [25], [26] utilize the trajectories of the sensor carrier or moving objects in the scene, together with geometric constraints, to estimate relative poses among sensors. These methods generally perform well in scenarios with stable and observable motion. However, if the scene is highly dynamic or if the motion trajectory is not controllable, their accuracy and robustness can degrade significantly.

Scene-based methods extract environmental features (e.g., edges, planes) or apply ICP-style iterative closest point algorithms [27]–[29] to register multi-end point clouds. These approaches do not require manual placement of calibration objects. Nevertheless, when initial values is inaccurate or in highly dynamic environments, these algorithms may fail to register or may suffer from reduced accuracy [30], [31]. To address these challenges some researchers have introduced global registration [32]–[34] or graph-based optimization frameworks [35]–[38] to reduce dependence on initial values and overlap areas. However, most of these algorithms still demand considerable computational resources for real-time performance.

C. Learning-Based Methods

In recent years, deep learning has made remarkable advances in the extraction and matching of 3D features, enabling end-to-end registration [39], [40]. Compared with classical geometric methods, deep learning can learn feature representations that are more robust to noise and dynamic interference using large-scale data. Nonetheless, these approaches typically require extensive labeled datasets or self-supervised signals, and the training process is time-consuming. Furthermore, domain adaptation issues may arise when applying the trained models to new scenarios—such as different urban structures or LiDARs with varying numbers of beams [41], [42].

D. Multi-end Spatial Registration Methods for Urban Intersection Scenarios

Urban intersections are challenging environments, often characterized by signal occlusions from surrounding buildings, unreliable GPS signals, and a high density of traffic participants with complex motion patterns. Multi-end spatial registration methods that rely on extrinsic priors or GPS positioning [6], [7], [11]–[14] are prone to failure in such settings. Meanwhile, classical ICP or feature-based registration

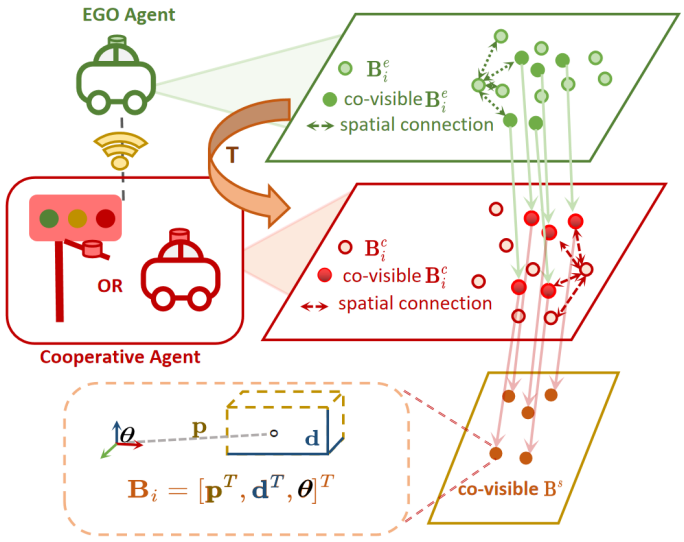


Fig. 3: Diagram of notations in problem formulation.

methods can perform poorly due to limited overlap and excessive noise. Although deep learning algorithms offer potential for tackling complex environments, they still face challenges related to data collection, model generalization, and computational efficiency in real-world applications. Additionally, when registering sensors with different configurations (such as beam counts, fields of view, resolutions, or scan frequencies), discrepancies in data resolution and feature representation must be carefully addressed [41]–[43].

In summary, developing a spatial registration method for multi-end sensing systems in urban intersections that is real-time and robust to initial pose values remains an open challenge.

III. PROBLEM FORMULATION

In this study, we construct a network of agents consisting of an Ego Agent and a Cooperative Agent (e.g., connected vehicles or roadside units), denoted as $\mathbf{A} = \{e, c\}$. We refer to the objects detected by these agents as passive objects, which do not participate in the information exchange. These objects serve as scene features, enriching the environmental description. Each object in the set of passive objects $\mathbf{B}$ perceived in the environment is represented as $\mathbf{B}_i = [\mathbf{p}^T, \mathbf{d}^T, \theta]^T \in \mathbb{R}^7$, where $\mathbf{p} \in \mathbb{R}^3$ denotes the center position, $\theta \in [0, 2\pi)$ represents the orientation, and $\mathbf{d} \in \mathbb{R}^3$ represents the 3D dimensions. We can also express $\mathbf{B}_i$ as a vertex matrix $\hat{\mathbf{B}}_i$. Here, $\hat{\mathbf{B}}_i \in \mathbb{R}^{3\times 8}$ represents the eight-vertex bounding box of the passive object, where each column corresponds to a 3D coordinate of a vertex. For convenience, the two representations $\mathbf{B}_i$ (emphasizing object aggregation) and $\hat{\mathbf{B}}_i$ (highlighting matrix-based transformations) will be used interchangeably to describe these passive objects in subsequent contexts.

From the perspectives of the Ego Agent and Cooperative Agent, the perceived object sets $\mathbf{B}^e$ and $\mathbf{B}^c$ are obtained, and their shared perceived objects are represented as $\mathbf{B}^s = \mathbf{B}^e \cap \mathbf{B}^c$.

It is important to note that $\mathbf{B}^e$ and $\mathbf{B}^c$ are located in their respective sensor coordinate systems, denoted as ${}^E\mathbf{B}^e$ and ${}^C\mathbf{B}^c$. However, for simplicity, we only label cross-coordinate sets, such as ${}^C\mathbf{B}^e$. Elements in $\mathbf{B}^s$ have been transformed into a unified coordinate system and will not be explicitly discussed further.

Previous studies typically rely on positioning systems to obtain a prior extrinsic value, primarily focusing on solving the extrinsic optimization problem. In this work, we aim to eliminate the need for prior extrinsic value by fully exploiting relationships between $\mathbf{B}^e$ and $\mathbf{B}^c$.

Then, a feature point cloud $\mathbf{P}$ is constructed from the object set $\mathbf{B}^s$ as follows:

$$\mathbf{P} = \left[\hat{\mathbf{B}}_1^s, \hat{\mathbf{B}}_2^s, \cdots, \hat{\mathbf{B}}_n^s\right]^\top, \quad \mathbf{P} \in \mathbb{R}^{8n\times 3} \tag{1}$$

where $\hat{\mathbf{B}}_i^s$ represents the i-th target in $\hat{\mathbf{B}}^s$, n denotes the number of co-viewing boxes.

The ultimate goal is to obtain the optimal extrinsic parameters $\hat{\mathbf{T}}$ by minimizing the distance between the corresponding feature point clouds:

$$(\hat{\mathbf{R}}, \hat{\mathbf{t}}) = \arg\min_{\mathbf{R},\mathbf{t}} \mathbf{E}(\mathbf{P}^e, \mathbf{R}\mathbf{P}^c + \mathbf{t}) \tag{2}$$

$$\hat{\mathbf{T}} = \begin{bmatrix} \hat{\mathbf{R}} & \hat{\mathbf{t}} \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{3}$$

where $\hat{\mathbf{R}} \in SO(3)$ is the rotation matrix, and $\hat{\mathbf{t}} \in \mathbb{R}^3$ is the translation vector. Together, they form the extrinsic parameters $\hat{\mathbf{T}} \in SE(3)$ as shown in Eq.(3). Eq.(2) performs the rotation and translation, transforming both point clouds into the same coordinate system. The error metric, $\mathbf{E}(\cdot)$, is computed using the L2 norm (Euclidean distance) between the feature point clouds, as detailed in Eq.(14).

IV. METHODOLOGY

A. Overview

The proposed V2X-Reg++ framework addresses the multi-LiDARs registration problem in V2X scenarios from a sensor perspective, while fundamentally solving a cross-view cross-source point cloud global registration problem at the data level. Instead of operating on dense, potentially noisy point clouds, which is computationally intensive and sensitive to outliers especially without initial alignment, our method operates at the semantic level of 3D detected objects. This stems from an intuitive hypothesis: *proper alignment of multi-view co-visible objects implies valid extrinsic transformations between perspectives*. This is particularly relevant for V2X urban intersections, which typically provide sufficient co-visible traffic participants (targets) to constrain the problem. Thus, we recast global registration primarily as a robust co-visible object matching problem, addressed in two main stages.

First, shared targets across agents are associated through spatial graph formed by targets, as illustrated in Fig. 4, where nodes denote objects and edges encode relative geometric constraints. This multi-end association mechanism is detailed in Section IV-B.

Subsequently, we convert matched targets into weighted feature point clouds $\mathbf{P}^e$ and $\mathbf{P}^c$ (Eq. 1), where confidence scores $\mathbf{M}_{i,j}$ (Eq. 11) guide a robust weighted SVD algorithm to solve extrinsic parameters, as elaborated in Section IV-C.

This dual-phase approach ensures registration accuracy by prioritizing high-confidence matches while adaptively suppressing noisy detections through graph-derived geometric constraints, achieving real-time performance without prior pose initialization.

B. Multi-End Object Association

Unlike existing object association methods that predominantly focus on short-term temporal continuity in object tracking scenarios [44], our approach emphasizes scene-level co-visible object matching under large perspective variations, as illustrated in Algorithm 1. We propose a dual-stage combinatorial strategy integrating singular value decomposition (SVD) and optimal transport (OT), with two main part: 1) **Scene-level correspondence mapping**: Transform potential pairwise object matches $(\mathbf{B}_i^e, \mathbf{B}_j^c)$ into coordinate-unified spatial graphs (Fig. 4) through coordinate alignment (Eq. 4); 2) **Optimal transport-based matching**: Convert the spatial graph association problem into a first-order node transportation task, where the *oDist* metric quantifies transportation costs by jointly evaluating match quantity ${}_\tau\overline{C}_{c_j}^{e_i}$ (Eq. 9) and precision ${}_\tau\overline{D}_{c_j}^{e_i}$ (Eq. 10), enabling robust object correspondence through optimal transport theory.

Algorithm 1 Multi-end Object Association

Input: Objects $\mathbf{B}^e$ and $\mathbf{B}^c$ from Ego and Cooperative Agents
Output: Matched object pairs $\mathbf{B}^s$
Initialize affinity matrix $\mathbf{M}$
for $\mathbf{B}_i^e$ in $\mathbf{B}^e$ **do**
  for $\mathbf{B}_j^c$ in $\mathbf{B}^c$ **do**
    Calculate $\mathbf{F}_j^i(\cdot)$ using Eq.(4)
    ${}^{E_{i,j}}\mathbf{B}^c = \mathbf{F}_j^i(\mathbf{B}^c)$ ▷ Transforming $\mathbf{B}^c$ based on localized transformation hypothesis $(\mathbf{B}_i^e, \mathbf{B}_j^c)$
    Calculate ${}_\tau\overline{D}_{c_j}^{e_i}$ and ${}_\tau\overline{C}_{c_j}^{e_i}$ using Eq.(9) and Eq.(10)
    if ${}_\tau\overline{D}_{c_j}^{e_i} < \tau_1$ **then**
      Update $\mathbf{M}_{i,j} = {}_\tau\overline{C}_{c_j}^{e_i}$
    else
      Update $\mathbf{M}_{i,j} = 0$
    end if
  end for
end for
Get assignment matrix $\mathbf{X}$ by solving Eq.(12) using [45]
Extract matched pairs $\mathbf{B}^S$ from $\mathbf{B}^e$ and $\mathbf{B}^c$ using $\mathbf{X}$
Return $\mathbf{B}^S$

1) Correspondence Mapping: Define $\mathbf{B}_i^e$ and $\mathbf{B}_j^c$ as the i-th and j-th perception objects in $\mathbf{B}^e$ and $\mathbf{B}^c$, respectively. The first step in our association pipeline is to generate hypothetical alignments. For every potential pair of objects $(\mathbf{B}_i^e, \mathbf{B}_j^c)$, we compute a relative transformation $\mathbf{F}_j^i(\cdot)$:

$$\mathbf{F}_j^i(\cdot) = \mathbf{R}_j^i \cdot + \mathbf{p}_i^e - \mathbf{R}_j^i \mathbf{p}_j^c \tag{4}$$

where $\cdot$ is a placeholder for any input point set, $\mathbf{p}_i^e$ and $\mathbf{p}_j^c$ are center position of $\mathbf{B}^e$ and $\mathbf{B}^c$ respectively. $\mathbf{R}_j^i$ is calculated as:

$$\mathbf{R}_j^i = \mathbf{U}\mathrm{diag}(1, 1, \det(\mathbf{U}\mathbf{V}^T))\mathbf{V}^T \tag{5}$$

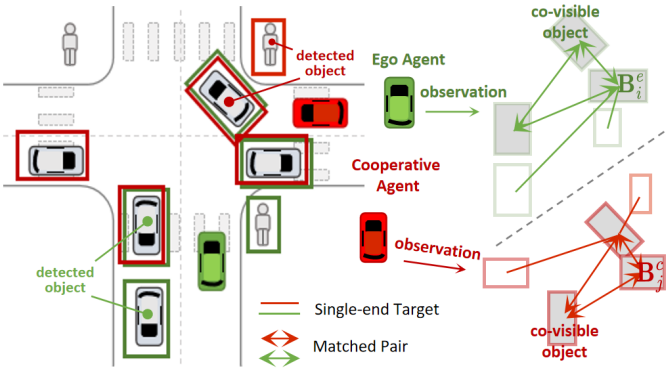


Fig. 4: The core idea of object association is to transform the correspondence between targets into a correspondence of the spatial graph formed by the objects.

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \hat{\mathbf{B}}_i^e(\hat{\mathbf{B}}_j^c)^T \tag{6}$$

Here, $\hat{\mathbf{B}}_i^e$ and $\hat{\mathbf{B}}_j^c$ are the vertex matrix representations of the objects $\mathbf{B}_i^e$ and $\mathbf{B}_j^c$, respectively, as defined in Section III. The matrices $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular matrices obtained by applying Singular Value Decomposition (SVD) to $\mathbf{H}$, following the method in [46].

Crucially, this initial SVD is not intended to yield the final extrinsic parameters. Instead, its purpose is to provide a localized transformation hypothesis. This hypothesis allows us to globally align the entire Cooperative Agent's perceived scene $\mathbf{B}^c$ with the Ego Agent's scene $\mathbf{B}^e$ if this specific object pair $(\mathbf{B}_i^e, \mathbf{B}_j^c)$ were a correct match. This potential localized alignment is fundamental for the subsequent Spatial Similarity Assessment.

2) Spatial Similarity Assessment: After generating a hypothetical alignment from a candidate object pair $(\mathbf{B}_i^e, \mathbf{B}_j^c)$, we introduce our novel *oDist* metric to assess its quality by quantifying the resulting scene-level spatial consistency.

To compute *oDist*, let ${}^{E_{i,j}}\mathbf{B}^c = \mathbf{F}_j^i(\mathbf{B}^c)$ denote the transformation of all objects $\mathbf{B}^c$ from the Cooperative Agent to the Ego Agent's coordinate system, based on the alignment hypothesis derived from the specific pair $(\mathbf{B}_i^e, \mathbf{B}_j^c)$. Within this hypothetically aligned scene, we define a valid matching pair set ${}_\tau\mathcal{V}_{c_j}^{e_i}$. This set comprises pairs of objects—one from $\mathbf{B}^e$ and one from the transformed ${}^{Ei,j}\mathbf{B}^c$—that satisfy a distance threshold τ. The specific mathematical definition is given as:

$${}_\tau\mathcal{V}_{c_j}^{e_i} = \{(\mathbf{B}_{i'}^e, {}^{E_{i,j}}\mathbf{B}_{j'}^c) \mid d(\mathbf{B}_{i'}^e, {}^{E_{i,j}}\mathbf{B}_{j'}^c) \leq \tau\} \tag{7}$$

$$d(\mathbf{B}_{i'}^e, {}^{E_{i,j}}\mathbf{B}_{j'}^c) = \alpha\|\mathbf{p}_{i'}^e - {}^{E_{i,j}}\mathbf{p}_{j'}^c\| + \beta\|\hat{\mathbf{B}}_{i'}^e - {}^{E_{i,j}}\hat{\mathbf{B}}_{j'}^c\| \tag{8}$$

Here, τ is a threshold designed to encompass potential matches, which can be empirically adjusted between 0 and 3 depending on the level of noise present in the scene. $\mathbf{p}_{i'}^e$ represents the spatial center of the object $\mathbf{B}_{i'}^e$, and $\hat{\mathbf{B}}_{i'}^e$ represents the vertex matrix of the i'-th perception object. α and β are weight factors that adjust the contribution of the location center and vertex distance of the detection frame to the index.

The *oDist* itself consists of two key components, both calculated from the valid matching set ${}_\tau\mathcal{V}_{c_j}^{e_i}$:

A confidence metric, ${}_{\tau}\overline{C}_{c_j}^{e_i}$, which measures how many other object pairs become well-aligned (i.e., fall within ${}_{\tau}\mathcal{V}_{c_j}^{e_i}$) under the current transformation hypothesis. This directly indicates the degree of matching for the candidate pair $(\mathbf{B}_i^e, \mathbf{B}_j^c)$.

$$ {}_{\tau}\overline{C}_{c_j}^{e_i} \doteq card({}_{\tau}\mathcal{V}_{c_j}^{e_i}) \tag{9} $$

A distance metric, ${}_{\tau}\overline{D}_{c_j}^{e_i}$, which quantifies how close the geometric correspondence is for these well-aligned pairs by calculating their average distance. This provides an indication of the potential matching error for the candidate pair $(\mathbf{B}_i^e, \mathbf{B}_j^c)$.

$$ {}_{\tau}\overline{D}_{c_j}^{e_i} \doteq \frac{\sum_{(\mathbf{B}_m, \mathbf{B}_n) \in {}_{\tau}\mathcal{V}_{c_j}^{e_i}} d(\mathbf{B}_m, \mathbf{B}_n)}{card({}_{\tau}\mathcal{V}_{c_j}^{e_i})} \tag{10} $$

In these equations, card$(\cdot)$ denotes the number of elements in a set, and $d(\cdot, \cdot)$ is defined in Eq.(8).

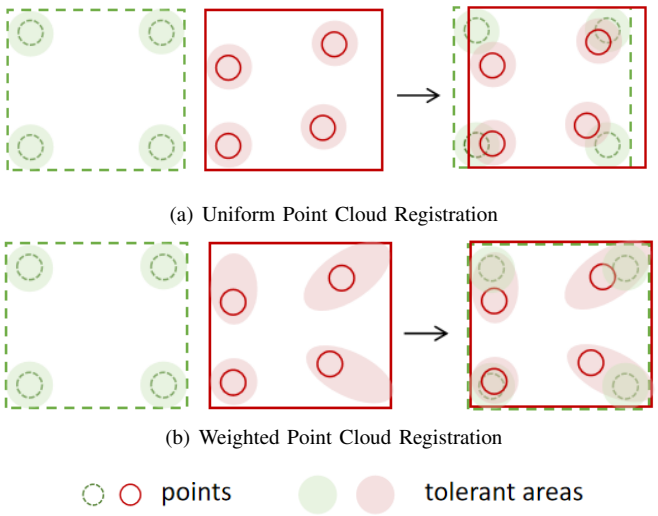
Therefore, the *oDist*, through its confidence ${}_{\tau}\overline{C}_{c_j}^{e_i}$ and distance ${}_{\tau}\overline{D}_{c_j}^{e_i}$ components, provides a rich, context-aware score. This score reflects the global plausibility of the initial local match $(\mathbf{B}_i^e, \mathbf{B}_j^c)$ by considering its impact on the entire scene's consistency, moving significantly beyond simple pairwise object similarity. This comprehensive assessment is crucial for robustly identifying true correspondences in the subsequent object matching stage.

3) Object Matching: The robust *oDist* scores form the basis of our object matching stage. Then, we construct an affinity matrix $\mathbf{M}$ between the perceived objects from the Ego Agent $\mathbf{B}^e$ and the Cooperative Agent $\mathbf{B}^c$. The elements $\mathbf{M}_{i,j}$ of this matrix quantify the likelihood of $\mathbf{B}_i^e$ and $\mathbf{B}_j^c$ being a correct match, based on the *oDist* components:

$$ \mathbf{M}_{i,j} = \begin{cases} {}_{\tau}\overline{C}_{c_j}^{e_i} & \text{if } {}_{\tau}\overline{D}_{c_j}^{e_i} < \tau_1, \\ 0 & \text{otherwise.} \end{cases} \tag{11} $$

where τ_1 represents a derived secondary filtering threshold, empirically adjustable between 0 and 2 based on the scene's noise level. This threshold applies to ${}_{\tau}\overline{D}_{c_j}^{e_i}$, which is the average distance of all valid matching pairs in ${}_{\tau}\mathcal{V}_{c_j}^{e_i}$. This is subtly different from τ in Eq.(7), which refers to the distance between single potential matching pairs. The threshold τ has a slightly larger range to include as many potential matching pairs as possible, whereas τ_1 is used to minimize the impact of local spatial graph similarities on the calculation of valid matching pair similarity. This affinity matrix $\mathbf{M}$ effectively captures the scene-level consistency for each potential object pair, as evaluated by *oDist*.

On this basis, we formulate an Optimal Transport (OT) problem to minimize the transformation cost from $\mathbf{B}^c$ to $\mathbf{B}^e$ based on *oDist*. Considering constraints in real scenarios, we set two restrictions: 1) Each target in $\mathbf{B}^c$ can be matched with at most one counterpart in $\mathbf{B}^e$; 2) Not all targets can find matches due to differences in perception conditions such as field of view and occlusions.



(a) Uniform Point Cloud Registration

(b) Weighted Point Cloud Registration

Fig. 5: Comparative illustration of uniform and weighted point cloud registration. In weighted point cloud registration, points with lower confidence have a larger range of spatial tolerant areas, allowing points with higher confidence to predominantly influence the registration outcome.

Based on the above, we define the shared target matching task as follows:

$$ \begin{aligned} &\underset{\mathbf{X}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m} -\mathbf{X}_{i,j} \mathbf{M}_{i,j} \qquad \text{s.t.} \quad \mathbf{X} \in \mathbf{\Pi} \\ &\mathbf{\Pi} = \{\mathbf{X} \mid \mathbf{X} \in \{0,1\}^{n_1 \times n_2}, \mathbf{X}\mathbf{1}_{n_2} \leq \mathbf{1}_{n_1}, \mathbf{X}^T\mathbf{1}_{n_1} \leq \mathbf{1}_{n_2}\} \end{aligned} \tag{12} $$

where n_1 and n_2 represent the number of elements in $\mathbf{B}^e$ and $\mathbf{B}^c$ respectively, and $\mathbf{X}$ is the assignment matrix. If $\mathbf{X}_{i,j} = 1$, it indicates that $\mathbf{B}_i^e$ and $\mathbf{B}_j^c$ are a matching pair, with the matching confidence given by $\mathbf{M}_{i,j}$. This constitutes a typical linear programming task, solved using [45]. The final set of shared objects $\mathbf{B}^s$ can be represented as:

$$ \mathbf{B}^s = \left\{ (\mathbf{B}_i^e, \mathbf{B}_j^c, \mathbf{M}_{i,j}) \mid {}_{i=1,\ldots,n_1; j=1,\ldots,n_2}^{\mathbf{X}_{i,j}=1} \right\} \tag{13} $$

C. Extrinsic Parameter Estimation

Once the set of shared object pairs $\mathbf{B}^s$ is established with associated matching confidences $\mathbf{M}_{i,j}$, we proceed to estimate the final extrinsic parameters. The corners of these matched 3D detection boxes are used to form feature point clouds $\mathbf{P}^e$ (for the Ego Agent) and $\mathbf{P}^c$ (for the Cooperative Agent), as per Eq.(1). Crucially, these are not uniform point clouds; each point pair derived from an object pair inherits the matching confidence $\mathbf{M}_{i,j}$ established in the object association phase, effectively making $\mathbf{P}^e$ and $\mathbf{P}^c$ weighted feature point clouds. As illustrated in Fig. 5, this weighting provides greater tolerance for points from less certain object matches, allowing the algorithm to prioritize the alignment of points from high-confidence object pairs.

The weighted point cloud registration problem can be formulated as follows:

$$(\hat{\mathbf{R}}, \hat{\mathbf{t}}) = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{\mathbf{p}_i^e \in \mathbf{P}^e, \mathbf{p}_i^c \in \mathbf{P}^c} w_i \|\mathbf{R}\mathbf{p}_i^e + \mathbf{t} - \mathbf{p}_i^c\|^2 \quad (14)$$

where $\mathbf{p}_i^e \in \mathbb{R}^3$ and $\mathbf{p}_i^c \in \mathbb{R}^3$ represent matched points in $\mathbf{P}^e$ and $\mathbf{P}^c$, respectively. The weight w_i is the matching confidence $\mathbf{M}_{i,j}$ from Eq.(11). $\hat{\mathbf{R}} \in SO(3)$ and $\hat{\mathbf{t}} \in \mathbb{R}^3$ represent the rotation matrix and translation vector of the target extrinsic parameters $\hat{\mathbf{T}}$ in Eq.(3).

To solve this problem, we construct the weighted cross-covariance matrix as follows:

$$\overline{\mathbf{H}} = \sum_{i=1}^{n} w_i (\mathbf{p}_i^e - \overline{\mathbf{p}}_e)(\mathbf{p}_i^c - \overline{\mathbf{p}}_c)^T \quad (15)$$

$$\overline{\mathbf{p}}_e = \frac{\sum_{i=1}^{n} w_i \mathbf{p}_i^e}{\sum_{i=1}^{n} w_i}, \quad \overline{\mathbf{p}}_c = \frac{\sum_{i=1}^{m} w_i \mathbf{p}_i^c}{\sum_{i=1}^{m} w_i} \quad (16)$$

where $\overline{\mathbf{p}}_e$ and $\overline{\mathbf{p}}_c$ are the weighted centers of $\mathbf{P}^e$ and $\mathbf{P}^c$, respectively.

Subsequently, singular value decomposition (SVD) [46] is applied to $\overline{\mathbf{H}}$ to obtain the left singular matrix $\mathbf{U}$ and right singular matrix $\mathbf{V}$, and these are used in Eq.(5) to determine the rotation matrix $\hat{\mathbf{R}}$. The translation vector $\hat{\mathbf{t}}$ can be obtained by:

$$\hat{\mathbf{t}} = \mathbf{P}_e - \hat{\mathbf{R}}\mathbf{P}_c \quad (17)$$

The final extrinsic parameters $\hat{\mathbf{T}}$ can be obtained using Eq.(3).

V. EXPERIMENT

This section details the experimental setup used to evaluate our method, including the metrics, datasets, and validation procedures.

A. Evaluation Metrics

To quantitatively evaluate our registration method, we first define the fundamental error metrics for a single trial: the Relative Rotation Error (*RRE*) and the Relative Translation Error (*RTE*). Based on these, we establish three key performance indicators that will be used for our final analysis: the Success Rate at a given threshold ($SuccessRate@\lambda$), the Mean Relative Rotation Error (*mRRE*@λ), and the Mean Relative Translation Error (*mRTE*@λ).

Relative Rotation Error (*RRE*): Measures the accuracy of the rotational part of the registration result, i.e., the angular difference between the estimated rotation matrix $\mathbf{R}_e$ and the true rotation matrix $\mathbf{R}_t$.

$$RRE = \arccos\left(\frac{\mathrm{tr}(\mathbf{R_t}^{-1}\mathbf{R_e}) - 1}{2}\right) \quad (18)$$

Relative Translation Error (*RTE*): Assesses the accuracy of the translation vector in the registration result, i.e., the distance difference between the estimated translation vector $\mathbf{t}_e$ and the true translation vector $\mathbf{t}_t$.

$$RTE = ||\mathbf{t}_t^{-1} - \mathbf{t}_e||_2 \quad (19)$$

$SuccessRate@\lambda$: The $SuccessRate@\lambda$ is defined as the proportion of registration trials in a total sample set $\mathcal{S} = \{1, 2, ..., N\}$ for which the achieved *RTE* is below a predefined threshold λ. This metric reflects the method's reliability in achieving a specified level of accuracy. The $SuccessRate@\lambda$ is mathematically expressed as:

$$\mathcal{S}_{\text{valid}}^{\lambda} = \{i \in \mathcal{S} | RTE_i < \lambda\} \quad (20)$$

$$SuccessRate@\lambda = \frac{|\mathcal{S}_{\text{valid}}^{\lambda}|}{|\mathcal{S}|} \quad (21)$$

Since $SuccessRate@\lambda$ only measures success frequency, we introduce the mean Relative Rotation Error (*mRRE*@λ) and mean Relative Translation Error (*mRTE*@λ) to quantify the accuracy of these successful trials. These metrics are calculated by averaging the *RRE* and *RTE* values exclusively over the subset of trials deemed successful by the threshold λ. The metrics are then defined as:

$$mRRE@\lambda = \frac{\sum_{i \in \mathcal{S}_{\text{valid}}^{\lambda}} \text{RRE}_i}{|\mathcal{S}_{\text{valid}}^{\lambda}|} \quad (22)$$

$$mRTE@\lambda = \frac{\sum_{i \in \mathcal{S}_{\text{valid}}^{\lambda}} \text{RTE}_i}{|\mathcal{S}_{\text{valid}}^{\lambda}|} \quad (23)$$

The choice of the threshold λ is critical and is guided by the operational requirements of Vehicle-to-Everything (V2X) applications, particularly for downstream tasks like cooperative perception and data fusion [6], [7], [11], [14], [47]. Drawing from existing literature and V2X system considerations (e.g., [14] suggests $\lambda = 2$m, and [47] suggests $\lambda = 3$m), and with recommendations from some perception algorithms [6], [7], [11], an *RTE* threshold λ in the range of 1 to 3 meters is often considered an acceptable upper bound for maintaining the requisite accuracy of these downstream tasks.

B. Dataset

In this study, we used the simulated dataset V2X-Sim [48] and the real-world dataset DAIR-V2X [49] for experimental validation. Both datasets contain extensive data collected from Vehicle-Everything Cooperative Autonomous Driving (VX-CAD) scenarios, including LiDAR data from vehicles and infrastructure as well as their 3D bounding box annotations and ground truth extrinsic parameters. The specifications of LiDAR Equipment are presented in Table I.

TABLE I: Specifications of LiDAR Equipment

Parameter	DAIR-V2X		V2X-Sim	
	R	V	R	V
LiDAR Points (lines)	300	40	32	32
Horizontal Field of View (°)	100	360	360	360
Max Detection Range (m)	280	200	70	70
Volume (frames)	3737	3737	1000	1000

Note: R = Roadside, V = Vehicle-side

One fundamental assumption for the effectiveness of the evaluation metrics discussed in the previous section is that the ground truth extrinsic parameters are sufficiently accurate.

To justify this assumption, we analyze the processing workflows of the two datasets. V2X-Sim [48] utilizes the SUMO [50] and CARLA [51] simulation platforms, which model the interaction between multiple vehicles and road-side units (RSUs) and the data acquisition process of their sensors. In this simulated environment, the data collection is considered to be free of delays and annotation errors, making the ground truth extrinsic parameters theoretically error-free. On the other hand, DAIR-V2X [49] is based on real-world data collected from actual scenarios. After filtering, cleaning, localization system transformation, and pose refinement, relatively accurate ground truth extrinsic parameters are obtained. Although some errors still exist, they are generally considered to be within an acceptable range.

C. Validation of Method Effectiveness

In this section, we perform comprehensive experimental validations of our method using the datasets and metrics outlined above. This includes verifying the theoretical validity and noise sensitivity of the method on simulated datasets and evaluating its performance under real noise conditions on real-world datasets. All experiments were conducted on a device with an Intel i7-9750H CPU.

1) Validity Verification on Simulated Dataset: We first validate the geometrical feasibility of our method under these perfect conditions. The specific results are shown in Table II.

TABLE II: Comparative Results on V2X-Sim.

Method	*mRRE*(°)	*mRTE*(m)	*SuccessRate*(%)		Time (s)
	@3°	@3m	@1m	@2m	
FGR [34]	0.69	0.16	78.64	95.15	0.92
Quartro [52]	0.17	0.18	96.40	98.20	0.83
Teaser++ [32]	0.77	0.17	76.70	94.17	0.91
V2X-Reg [23]	0.06	0.03	93.26	95.48	0.37
V2X-Reg++	**0.01**	**0.01**	**96.80**	**98.31**	**0.13**

We find that under conditions of perfect detection and perfect transmission, V2X-Reg++ achieves near-zero rotation and translation deviations quickly, proving the geometrical correctness of the perception-based approach. It is also observed that even in ideal conditions, the success rate of the method is not 100%. This is due to the spatial information between perception objects occasionally having similarities, which can affect our method. However, these coincidental spatial similarities are rare, and the likelihood can be reduced by increasing the number of perception objects.

2) Sensitivity to Noise: To evaluate our method's robustness against realistic perception errors, we conducted a noise injection experiment on the V2X-Sim dataset [48]. We grounded our noise model in the performance of a widely-used 3D detector, PointPillars [53], on the nuScenes benchmark [54]. On this benchmark, PointPillars achieves a mean Average Translation Error (mATE) of $0.32m$ and a mean Average Orientation Error (mAOE) of $0.28rad$ ($\approx 16°$).

Recognizing that these are average metrics and that individual detections, especially in challenging scenarios, can exhibit larger deviations, our experiment was designed to span a wider spectrum. We injected Gaussian noise into the object's position

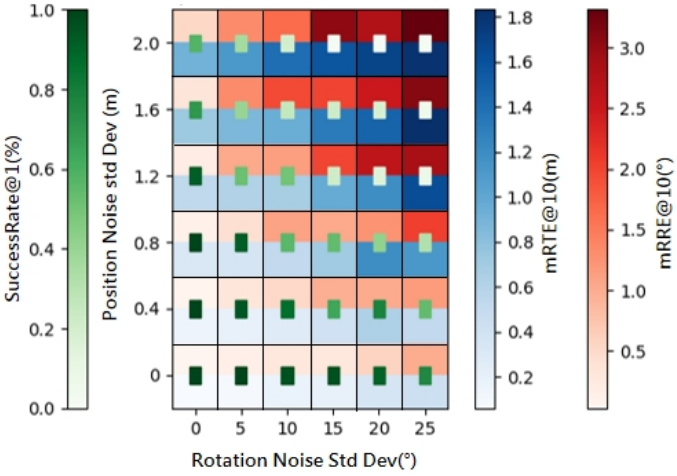


Fig. 6: Heatmap Analysis of three performance metrics under varying noise levels applied to ground truth bounding boxes in the V2X-Sim Dataset. Despite increasing errors with higher input noise, the method shows strong noise resilience.

(mean $\mu_1 = 0$ m, with standard deviation from 0 to 2.0m) and von Mises noise into its orientation (mean $\mu_2 = 0°$, with standard deviation from 0° to 25°). This range effectively covers performance from typical error levels to more severe failure cases.

As shown in Fig. 6, while the method is resilient to individual error dimensions, a combination of translation and rotation noise leads to a more rapid decline in effectiveness. Encouragingly, our method still demonstrates a strong corrective capability even under high noise. Specifically, the maximum *mRTE@10* and *mRRE@10* were controlled at $1.8m$ and 3.5°, respectively. This confirms the method's robustness across a realistic spectrum of perception noise.

3) Validity Verification on Real-World Dataset: To assess our method's practical efficacy, we performed experiments on the DAIR-V2X dataset [49]. Its varied LiDAR configurations (Table I) and sparse point cloud overlaps (Fig. 7) pose substantial registration challenges.

Performance and Accuracy.

Comparison with methods requiring initial values: The upper part of Table III contextualizes our results against methods reliant on initial poses. Notably, real-world V2X point clouds suffer from heterogeneity and cross-view discrepancies. Coupled with potential inaccuracies in DAIR-V2X ground-truth extrinsics, even initial-value-based methods [56], [58] under ideal conditions ($Noise = 0m\ and\ 0°$) exhibit an approximate 0.5m and $0.5°$ error. This can be seen as an accuracy *upper bound* for this dataset. Introducing equivalent rotational and translational noise to these methods rapidly degrades their accuracy, rendering them nearly unusable at $2m\ and\ 2°$ noise. V2X-Reg++, however, effectively mitigates initial pose deviations. Its performance remains comparable to the dataset accuracy upper bound (e.g., the *mRTE*@$1m$) and demonstrates stability against increasing initial errors, underscoring its practical value.

Comparison with initial-value-free methods: The lower part of Table III shows V2X-Reg++ significantly surpass-

TABLE III: Comparative Results on the DAIR-V2X Dataset. For the methods that require initial pose values, we add noise of equal magnitude to the rotational and translational dimensions to simulate different levels and sources of noise in real-world scenarios. Lower values are better for *mRRE* and *mRTE* (↓), and higher values are better for $SuccessRate$ (↑). Subscripts GT, PP, and SC denote ground-truth boxes, PointPillars [53] detector boxes, and SECOND [57] detector boxes, respectively. The superscript k signifies the use of top-k dimension-sorted boxes, while ∞ indicates use all boxes provided. The **best** and second-best results are highlighted in each section.

Init	Noise (m & °)	Method	*mRRE* (°) ↓			*mRTE* (m) ↓			$SuccessRate$ (%) ↑			Time (s) ↓
			@1°	@2°	@3°	@1m	@2m	@3m	@1m	@2m	@3m	
	0		0.65	0.98	1.07	**0.42**	**0.54**	**0.58**	47.52	89.55	96.01	2.91
	1	ICP [55]	0.80	1.36	1.72	0.66	1.31	1.62	0.86	37.93	80.50	2.92
	2		0.00	1.48	2.11	0.00	1.33	2.03	0.00	3.66	19.94	2.86
	0		**0.52**	**0.80**	**0.88**	**0.42**	**0.54**	0.57	**59.59**	**90.41**	96.12	1.35
	1	PICP [56]	0.74	1.31	1.67	0.75	1.32	1.63	2.91	42.78	87.93	1.76
✓	2		0.80	1.40	2.11	0.53	1.45	2.10	0.22	2.69	21.12	1.70
	0		0.63	0.89	0.99	0.54	0.78	0.89	54.20	88.69	**97.63**	0.46
	1	VIPS [14]	0.66	1.04	1.24	0.54	0.82	1.02	18.53	39.01	47.74	0.44
	2		0.58	1.17	1.56	0.48	0.96	1.39	2.37	7.87	13.15	0.47
	0		0.61	0.97	1.21	0.53	0.80	1.06	17.11	23.04	26.49	**0.35**
	1	CBM [12] †	0.71	0.94	1.14	0.61	0.74	1.00	9.91	15.63	16.49	0.36
	2		0.69	1.09	1.38	0.58	0.76	1.06	6.03	12.28	16.81	**0.35**
	-	FGR [34]	0.71	1.15	1.47	0.70	1.13	1.45	14.76	31.57	35.34	22.73
	-	Quartro [52]	**0.62**	1.22	1.46	0.65	1.19	1.51	12.07	30.50	45.04	21.58
	-	Teaser++ [32]	0.69	1.13	1.47	0.66	1.09	1.44	14.33	29.74	34.81	22.43
×	-	V2X-Reg [23]	0.66	1.03	1.25	0.54	0.91	1.18	25.54	55.93	72.31	0.21
	-	V2X-Reg++$_{GT}^{\infty}$	**0.62**	**1.01**	1.26	**0.49**	**0.83**	1.07	22.88	48.03	61.49	0.46
	-	V2X-Reg++$_{GT}^{25}$	0.63	**1.01**	**1.23**	0.52	0.85	**1.05**	**32.27**	**67.59**	**82.93**	0.12
	-	V2X-Reg++$_{GT}^{15}$	0.65	1.05	1.30	0.54	0.87	1.10	26.79	61.17	78.75	0.09
	-	V2X-Reg++$_{GT}^{10}$	0.66	1.11	1.36	0.57	0.92	1.15	20.02	54.86	71.98	**0.04**
	-	V2X-Reg++$_{PP}^{15}$	0.66	1.06	1.29	0.55	0.86	1.07	24.91	56.62	70.94	-
	-	V2X-Reg++$_{SC}^{15}$	0.65	1.05	1.29	0.54	0.86	1.06	25.15	56.89	71.23	-
	-	V2X-Reg++$_{GT}^{25}$(hSVD)‡	0.71	1.13	1.35	0.62	0.98	1.25	21.82	60.43	74.92	0.12
	-	V2X-Reg++$_{GT}^{25}$(mSVD)‡	0.67	1.08	1.31	0.56	0.94	1.19	25.22	63.58	80.22	0.12

† : For CBM [12], our reimplementation (due to partial code availability) achieves comparable accuracy but significantly lower success rates under $SuccessRate@\lambda$. We will make it open-source in our codebase.

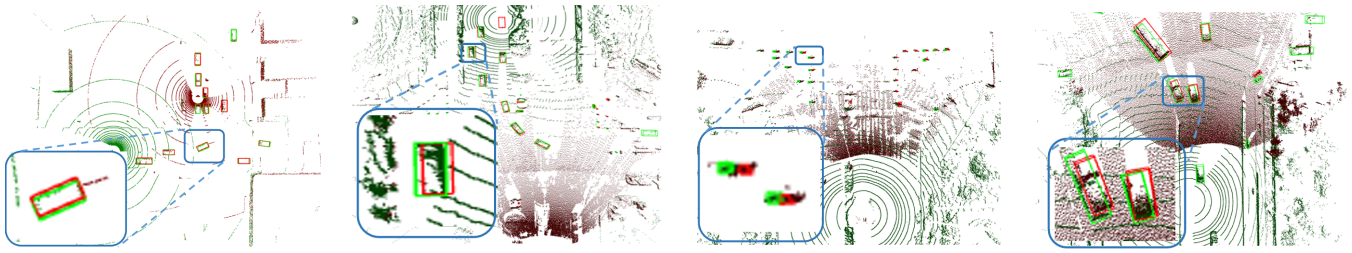
‡ : V2X-Reg++ entries without parentheses (e.g., V2X-Reg++$_{GT}^{25}$) use the proposed Weighted SVD (wSVD) by default. Comparisons between wSVD, mSVD, and hSVD strategies (Section V-E2) validate wSVD's superior robustness.

ing other initial-value-free techniques. It achieves a high $SuccessRate@2m$ of 56.89% and low *mRTE*@$2m$ of 0.86m, even with real-world detector outputs from SECOND [57] and PointPillars [53]. This suggests that employing detection boxes as an intermediate representation for global registration of cross-source point clouds with large FoV differences can match the accuracy of point-based or feature-based methods. Crucially, it also enables real-time processing and substantially reduces transmission bandwidth, making it well-suited for V2X challenges.

An interesting observation is that while point/feature-based methods [56], [58] achieve higher accuracy with initial values than detection-box-based methods [12], [14] , while ours are more accurate without initial values, outperforming methods like [32], [34], [52], indicating that the high-level semantic information provided by detection boxes is better suited for coarse registration tasks.

Impact of Detection Uncertainty. To quantify the impact of real-world perception noise, we used detections from PointPillars (PP) [53] and SECOND (SC) [57] instead of ground-truth (GT) boxes. While perception errors cause slight performance degradation, V2X-Reg++ remains robust overall. The impact of detector errors on overall registration accuracy in real scenes is considerably smaller than in simulations (Fig. 6), sometimes even yielding better *mRTE*@$3m$ values. The main adverse effect of detection errors is on $SuccessRate@\lambda$. We posit this is due to inherent spatial deviations in V2X scenarios, which dilute the influence of perceptual errors on extrinsic parameters in successful registrations. Stronger perceptual errors tend to cause registration failure and are thus excluded from accuracy statistics. The overall decrease in $SuccessRate@\lambda$ is acceptable, reflecting our method's robustness to perceptual noise by emphasizing high-confidence matches. In this context, perceptual uncertainty's impact can be mitigated by pre-processing, such as expanding detection ranges and filtering low-confidence boxes, enhancing practical viability.

Detection Box Quantity Effects. Our method's reliance on detection boxes means their quantity is a key factor, alongside perceptual uncertainty. DAIR-V2X frames exhibit wide variation in box counts (from a few to tens). Our SVD-based extrinsic search (Section IV-B1) naturally favors larger-dimension boxes due to greater spatial distinctiveness and potentially higher annotation/detection accuracy. We thus sorted boxes by volume and experimented with top-K subsets. Table III indicates that performance generally improves with more boxes. However, using all boxes (V2X-Reg++$_{GT}^{\infty}$) degrades performance. Fig. 7(c) illustrates this with small traffic cones from DAIR-V2X, which are challenging to annotate/detect accurately. Their low spatial distinctiveness leads to larger SVD-derived extrinsic errors, impacting overall performance. Thus, pre-filtering input boxes significantly boosts performance and



(a) Ideal Scenario from V2X-Sim (b) Ideal Scenario from DAIR-V2X (c) Challenging Scenario (d) Complex Scenario

Fig. 7: Comparative Registration Results Across Diverse Scenarios: (a, d) Accurate alignment of multi-end perception objects. (b) Registration errors due to small-sized perception objects. (d) Suboptimal registration results due to inherent perception errors.

reduces computation time. Comparing V2X-Reg++$_{\text{GT}}^{15}$, V2X-Reg++$_{\text{PP}}^{15}$, and V2X-Reg++$_{\text{GT}}^{10}$ reveals that box quantity has a greater impact on the method's performance than individual box uncertainty. This aligns with our method's focus on information derived from the set of boxes, making set size (quantity) more influential than individual box characteristics (uncertainty).

Runtime Analysis. Preliminary multi-threaded Python optimizations on an Intel i7-9750H CPU @2.6GHz yield a runtime of 0.09s for V2X-Reg++$_{\text{GT}}^{15}$, which is significantly below the $0.35s$ requirement in [13]. This result is primarily intended for relative comparison with existing methods, highlighting the algorithm's real-time capability. Furthermore, the approach exhibits substantial parallelization potential. Specifically, the computational bottleneck arises from the $O(N^2)$ *oDist* calculations for potential pairs; however, these operations are mutually independent and thus amenable to parallelization. Additionally, the underlying matrix operations are well-suited for GPU acceleration. While the runtime upper bound occurs with V2X-Reg++$_{\text{GT}}^{\infty}$(dozens of boxes), we note that such configurations may reduce accuracy, as discussed earlier. Therefore, practical deployment requires a trade-off between selected box quantity and runtime efficiency.

D. Effectiveness Test of Overall Distance

The multi-end spatial registration method proposed in this paper centers around the *oDist* metric for achieving target association. Compared to *oIoU* metric proposed in our previous method V2X-Reg [23], the *oDist* exhibits superior performance and can better assess associations with objects that are further away. To validate the trends of the *oDist* metric under various noise conditions, we designed experiments on DAIR-V2X [49].

As shown in Fig. 8, compared to the *oIoU* metric, the *oDist* proposed in this paper demonstrates a smoother curve of change with respect to different deviations in the external parameters. This implies that it is less susceptible to falling into local mathematical optima, thus providing better monitoring performance for external parameter alignment.

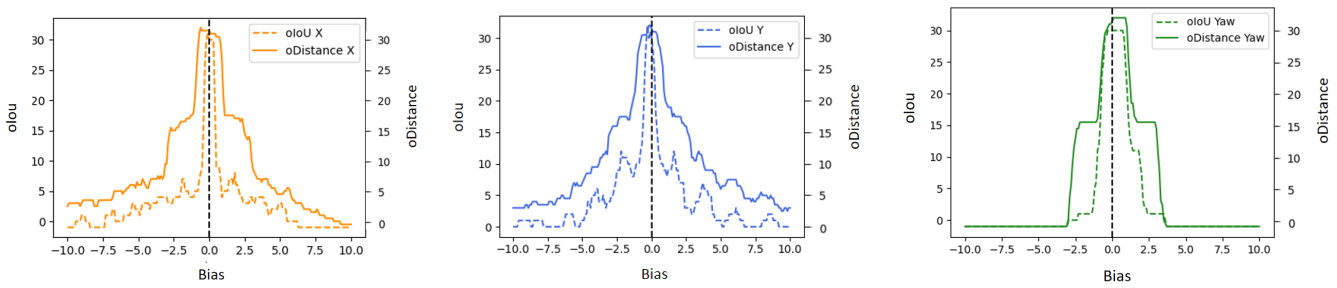
It is worth explaining that under continuously increasing biases, most perception-object-based metrics exhibit this type of local rise in values, due to the lack of confirmed associations between multi-end perception targets during the monitoring of external parameters. This causes an occasional rise in metrics when $\mathbf{B}_i^e$ moves away from $\mathbf{B}_j^c$ and closer to $\mathbf{B}_{j+1}^c$. In contrast to the *oIoU* metric, the *oDist* metric proposed in this paper extends the valid association phase from merely considering the *IoU* of two targets greater than zero to a set threshold distance determination, akin to the improvements suggested by softNMS [59] on the classical NMS algorithm.

E. Ablation Experiments

In this section, we aim to validate the efficiency of the object association module introduced in Section IV-B and the external parameter optimization module discussed in Section IV-C through a series of ablation experiments.

1) Comparison of Object Association Strategies: To validate the superiority of the target association module strategy proposed in this paper, we conducted a comparative analysis with the angle-based and length-based association strategies defined in [14], as well as the *oIoU* metric association strategy defined in [23]. As shown in Fig. 9, the proposed association strategy (Strategy 1) and the *oIoU* metric association strategy (Strategy 2) both achieved high target association rates. However, the *oDist* metric strategy excelled, achieving superior final success rates for both $\lambda = 1$ and $\lambda = 2$, thus confirming its robustness. The angle-based (Strategy 3) and length-based (Strategy 4) association strategies defined in [14], originally intended for scenarios with initial external parameter values, did not perform as well overall. Nonetheless, they demonstrated some applicability, especially the length-based association strategy 4, which showed certain performance characteristics without initial external values.

2) Comparison of Strategies for External Parameter Solution: To validate the superiority of the external parameter optimization module discussed in Section IV-C, we compared three strategies: the Weighted SVD (wSVD) method proposed in this paper, the Mean SVD (mSVD) method, and the SVD method based on the highest confidence detection box (hSVD). As shown at the bottom of Table. III, we observed that the Weighted SVD method proposed in this paper demonstrated enhanced robustness.



(a) Comparison of metrics along the x-axis (b) Comparison of metrics along the y-axis (c) Comparison of metrics along the yaw dimension

Fig. 8: Verification of the indicator effects of the *oDist* metric proposed in this paper and the *oIoU* metric proposed in [23] on the initial external parameters with added noise in different directions on DAIR-V2X. It is observed that the *oDist* metric proposed by this method displays smoother curves, indicating better performance of *oDist* in monitoring the alignment level of external parameters.

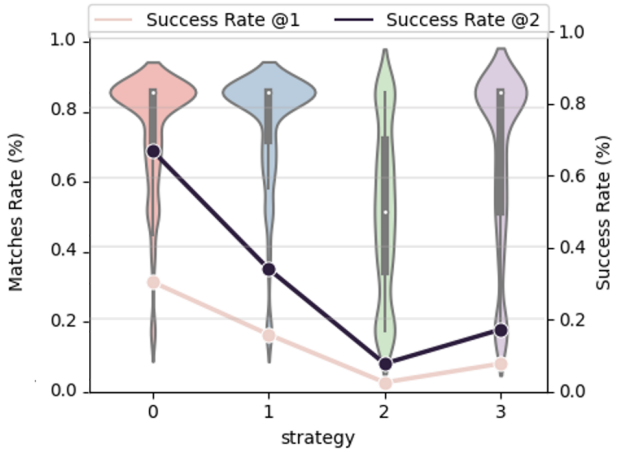


Fig. 9: Violin plot comparing the effects of different object association strategies. The Matches Rate refers to the ratio of the number of correctly matched objects to the number of ground truth matched objects annotated in the scene. Strategy 1 is the *oDist* metric proposed in this paper, Strategy 2 is the *oIoU* metric proposed by [23], and Strategies 3 and 4 are the angular and length similarity metrics commonly used in papers such as [14]. The effectiveness of the target association strategy adopted in this paper is evident.

VI. Practical Use Cases

To demonstrate the practical value of V2X-Reg++, this section outlines a typical scenario illustrating its role in enhancing the robustness of V2X systems from initial deployment, particularly focusing on boot-up safety registration and its synergy with positioning systems.

V2X-Reg++ can significantly enhance the safety and reliability of V2X systems during initial deployment and operation. Through the introduced *oDist* metric, the system can continuously monitor the spatial alignment quality of multi-end sensing system. If a registration deviation is detected at system boot-up due to poor initial extrinsic parameters (e.g., a bad *oDist* value), or if misalignment occurs during operation due to parameter drift, V2X-Reg++ will be triggered to quickly optimize the extrinsic parameters, as detailed in Algorithm 2. This mechanism creates a robust synergy with positioning systems. In environments like urban canyons where positioning signals are unreliable, V2X-Reg++ acts as a perception-level fail-safe, providing the precise and continuous alignment required for all V2X applications.

VII. Discussion

A. Multi-End Spatial Registration Tasks and Localization Tasks

While localization and multi-end spatial registration are related, as both express spatial relationships, they are fundamentally distinct tasks in terms of their objectives, target objects, and role in the autonomous driving pipeline. Specifically, localization typically aims to determine a vehicle's absolute position within a global coordinate system, focusing on the vehicle's body frame. In contrast, registration, the focus of this paper, seeks to resolve the relative pose transformation between different sensor coordinate systems. This places as a foundational upstream task, responsible for enabling the accurate spatial alignment of sensor data, whereas localization is a downstream application that often relies on this pre-aligned data. Furthermore, the real-time nature of our method should not be misconstrued as a characteristic exclusive to localization. While classic was often a static, offline process, modern targetless methods are increasingly evolving towards continuous, real-time operation to meet the demands of V2X systems. Therefore, based on its focus on relative sensor poses and its role as an upstream enabler for data fusion, our method firmly resides within the domain of sensor registration.

B. Trends in Multi-End Spatial Registration

The development of multi-sensor spatial registration technology, as the basis for multi-sensor data fusion, largely depends on the demands of subsequent perception fusion tasks. Similarly, the trends in multi-end spatial registration tasks are influenced by multi-end perception fusion tasks. One trend in the latter is to achieve better perception effects under inaccurate external parameters [6], [15], [60], [61]. From another perspective, this integrates the registration process

Algorithm 2 V2X-Reg++ Registration & Monitoring

Input:
  P_1, P_2: Detection boxes from two sensors
  θ_b: Boot threshold for initial registration
  θ_m: Monitor threshold for runtime checks
  R_{max}: Max retry attempts
Output: $\mathbf{T}_{cur}$: Current extrinsics

if stored $\mathbf{T}_{stor}$ exists **then**
  $\mathbf{T}_{cur} \leftarrow \mathbf{T}_{stor}$
else
  $\mathbf{T}_{cur} \leftarrow$ null
end if

if $\mathbf{T}_{cur} =$ null **or** $oDist(P_1, P_2, \mathbf{T}_{cur}) > \theta_b$ **then**
  $\mathbf{T}_{cur} \leftarrow$ Registration$(P_1, P_2, \theta_b, R_{max})$
  if $\mathbf{T}_{cur} =$ fail **then trigger alert**
  end if
end if

while system running **do**
  Acquire new P_1, P_2
  if $oDist(P_1, P_2, \mathbf{T}_{cur}) > \theta_m$ **then**
    $\mathbf{T}_{new} \leftarrow$ Registration$(P_1, P_2, \theta_m, R_{max})$
    if $\mathbf{T}_{new} \neq$ fail **then**
      $\mathbf{T}_{cur} \leftarrow \mathbf{T}_{new}$; **store** $\mathbf{T}_{cur}$
    else
      degrade operation
    end if
  end if
  apply $\mathbf{T}_{cur}$
end while

function REGISTRATION(P_1, P_2, θ, R)
  $r \leftarrow 0$
  while $r < R$ **do**
    $\mathbf{T} \leftarrow$ V2X-Reg++(P_1, P_2)
    if $oDist(P_1, P_2, \mathbf{T}) \leq \theta$ **then**
      return $\mathbf{T}$ ▷ Success
    end if
    $r \leftarrow r + 1$
  end while
  return fail ▷ All attempts failed
end function

into the perception task, no longer treating registration as an independent output but as a dynamically adjusted intermediate quantity within the perception algorithm. This integration of registration and perception aligns with the trend towards end-to-end development in autonomous driving. However, these methods still exhibit a low tolerance for deviations in external parameters. Under current research, multi-end spatial registration remains a necessary step.

VIII. CONCLUSIONS

The main contribution of V2X-Reg++ is that it overcomes the limitations of existing multi-end spatial registration methods that rely on positional priors, enabling effective registration of multi-end sensing systems in environments with unstable positioning signals, such as urban canyons. The method combines a two-stage SVD algorithms and optimal transport theory to effectively solve the problem of data consistency among multi-end sensors. Additionally, our proposed *Overall Distance* (*oDist*) metric offers a reliable method for monitoring the quality of the spatial alignment in real-time.

Extensive experiments on the V2X-Sim and DAIR-V2X datasets demonstrate the effectiveness and robustness of V2X-Reg++. It provides stable, high-precision registration where initial-value-dependent methods fail due to noise, and it outperforms other initial-value-free approaches in both success rate and computational efficiency. The method is also robust to perception noise, maintaining low registration errors even when using noisy detections from real-world detectors, as shown in our DAIR-V2X experiments. Its computational efficiency, with runtimes as low as 0.1 seconds, fully satisfies real-time demands. However, its primary limitation is a direct dependency on the quantity and quality of co-visible detection boxes. As performance can degrade with either too few objects or an excess of low-quality detections, this paper also proposes some pre-processing strategies in the experimental analysis section to mitigate these effects.

Future research will continue to explore the application potential of V2X-Reg++ in broader urban environments. This includes integrating other sensor types, leveraging static information from maps to extend its applicability to scenes with fewer dynamic objects, and scaling the algorithm for multi-agent systems. Expanding the current pairwise framework to robustly handle multiple agents in multi-frame asynchronous scenarios is a key direction, potentially by using our pairwise results as edges in a global pose-graph optimization network. However, this introduces communication bottlenecks, and future work will investigate co-optimizing registration with communication strategies. On the performance front, given the algorithm's reliance on matrix operations, significant speed-ups are anticipated from GPU-accelerated implementations. Additionally, as autonomous driving technologies advance, integrating V2X-Reg++ more deeply with the full autonomous driving stack will be a critical area for subsequent research. By providing a robust and real-time solution for initial-value-free sensor registration, we believe V2X-Reg++ offers a foundational technology essential for advancing cooperative perception and, consequently, improving the safety and intelligence of future automated urban traffic systems.

REFERENCES

[1] D.-J. Lin, M.-Y. Chen, H.-S. Chiang, and P. K. Sharma, "Intelligent traffic accident prediction model for internet of vehicles with deep learning approach," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 3, pp. 2340–2349, 2021.

[2] X. Huang, P. Lin, C. Chen, B. Ran, and M. Tan, "Dynamic trajectory-based traffic dispersion method for intersection traffic accidents in an intelligent and connected environment," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 1, pp. 84–100, 2021.

[3] L. Luo, L. Sheng, H. Yu, and G. Sun, "Intersection-based v2x routing via reinforcement learning in vehicular ad hoc networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5446–5459, 2021.

[4] M. Yang, B. Ai, R. He, Z. Ma, H. Mi, D. Fei, Z. Zhong, Y. Li, and J. Li, "Dynamic v2v channel measurement and modeling at street intersection scenarios," *IEEE Transactions on Antennas and Propagation*, vol. 71, no. 5, pp. 4417–4432, 2023.

[5] P. Sun, D. Nam, R. Jayakrishnan, and W. Jin, "An eco-driving algorithm based on vehicle to infrastructure (v2i) communications for signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 144, p. 103876, 2022.

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

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

[8] Y. Xiong, X. Zhang, X. Gao, Q. Qu, C. Duan, R. Wang, J. Liu, and J. Li, "Cooperative camera-lidar extrinsic calibration for vehicle-infrastructure systems in urban intersections," *IEEE Internet of Things Journal*, 2025.

[9] X. Zhang, Y. Xiong, Q. Qu, R. Wang, X. Gao, J. Liu, S. Guo, and J. Li, "Cooperative visual-lidar extrinsic calibration technology for intersection vehicle-infrastructure: A review," *arXiv preprint arXiv:2405.10132*, 2024.

[10] X. Zhang, Y. Xiong, Q. Qu, S. Zhu, S. Guo, D. Jin, G. Zhang, H. Ren, and J. Li, "Automated extrinsic calibration of multi-cameras and lidar," *IEEE Transactions on Instrumentation and Measurement*, 2023.

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

[12] Z. Song, T. Xie, H. Zhang, J. Liu, F. Wen, and J. Li, "A spatial calibration method for robust cooperative perception," *IEEE Robotics and Automation Letters*, 2024.

[13] Y. He, L. Ma, Z. Jiang, Y. Tang, and G. Xing, "Vi-eye: semantic-based 3d point cloud registration for infrastructure-assisted autonomous driving," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 573–586.

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

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

[16] P. Xie and M. G. Petovello, "Measuring gnss multipath distributions in urban canyon environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 366–377, 2014.

[17] M. Talas *et al.*, "Connected vehicle pilot deployment program phase 3 – system performance report: New york city," U.S. Department of Transportation, Tech. Rep. FHWA-JPO-18-715, December 2021. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/63102

[18] ——, "Connected vehicle pilot deployment program phase 3, understanding and enabling cooperative driving for advanced connected vehicles in new york city – new york city department of transportation (nycdot)," U.S. Department of Transportation, Tech. Rep. FHWA-JPO-21-920, December 2021, table 74. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/63613

[19] C. Sanders and Y. Wang, "Localizing spoofing attacks on vehicular gps using vehicle-to-vehicle communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 656–15 667, 2020.

[20] G. Twardokus and H. Rahbari, "Vehicle-to-nothing? securing c-v2x against protocol-aware dos attacks," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1629–1638.

[21] F. Wang, Y. Hong, and X. Ban, "Infrastructure-enabled gps spoofing detection and correction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13 878–13 892, 2023.

[22] X. Huang, G. Mei, and J. Zhang, "Cross-source point cloud registration: Challenges, progress and prospects," *Neurocomputing*, p. 126383, 2023.

[23] Q. Qu, Y. Xiong, X. Wu, H. Li, and S. Guo, "V2i-calib: A novel calibration approach for collaborative vehicle and infrastructure lidar systems," *arXiv preprint arXiv:2407.10195*, 2024.

[24] A. Dhall, K. Chelani, V. Radhakrishnan, and K. M. Krishna, "Lidar-camera calibration using 3d-3d point correspondences," *arXiv preprint arXiv:1705.09785*, 2017.

[25] J. Lv, J. Xu, K. Hu, Y. Liu, and X. Zuo, "Targetless calibration of lidar-imu system based on continuous-time batch estimation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9968–9975.

[26] Y. Xiong, X. Zhang, W. Gao, Y. Wang, J. Liu, Q. Qu, S. Guo, Y. Shen, and J. Li, "Gf-slam: a novel hybrid localization method incorporating global and arc features," *IEEE Transactions on Automation Science and Engineering*, 2024.

[27] X. Huang, G. Mei, J. Zhang, and R. Abbas, "A comprehensive survey on point cloud registration," *arXiv preprint arXiv:2103.02690*, 2021.

[28] J. Yang, H. Li, D. Campbell, and Y. Jia, "Go-icp: A globally optimal solution to 3d icp point-set registration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2241–2254, 2015.

[29] K. Koide, M. Yokozuka, S. Oishi, and A. Banno, "Voxelized gicp for fast and accurate 3d point cloud registration," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 054–11 059.

[30] Y. Zheng, Y. Li, S. Yang, and H. Lu, "Global-pbnet: A novel point cloud registration for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22 312–22 319, 2022.

[31] C. Shi, X. Chen, K. Huang, J. Xiao, H. Lu, and C. Stachniss, "Keypoint matching for point cloud registration using multiplex dynamic graph attention networks," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8221–8228, 2021.

[32] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[33] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu, "Geotransformer: Fast and robust point cloud registration with geometric transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9806–9821, 2023.

[34] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 766–782.

[35] 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*, 2021.

[36] C. Zhao, D. Ding, Y. Shi, Y. Ji, and Y. Du, "Graph matching-based spatiotemporal calibration of roadside sensors in cooperative vehicle-infrastructure systems," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[37] C. Li, L. Xu, C. Jin, and L. Wang, "Graphps: Graph pair sequences-based noisy-robust multi-hop collaborative perception," *IEEE Transactions on Intelligent Vehicles*, 2023.

[38] 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*, vol. 24, no. 5, pp. 5094–5104, 2023.

[39] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, "Pointnetlk: Robust & efficient point cloud registration using pointnet," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7163–7172.

[40] Y. Wang and J. M. Solomon, "Deep closest point: Learning representations for point cloud registration," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3523–3532.

[41] Y. Wu, H. Ding, M. Gong, A. K. Qin, W. Ma, Q. Miao, and K. C. Tan, "Evolutionary multiform optimization with two-stage bidirectional knowledge transfer strategy for point cloud registration," *IEEE Transactions on Evolutionary Computation*, vol. 28, no. 1, pp. 62–76, 2024.

[42] H. Ding, H. Xu, Y. Wu, H. Li, M. Gong, W. Ma, Q. Miao, J. Shi, and Y. Lei, "Evolutionary multitasking with two-level knowledge transfer for multi-view point cloud registration," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2024, pp. 304–312.

[43] X. Huang, G. Mei, and J. Zhang, "Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 366–11 374.

[44] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2021.

[45] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016.

[46] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on pattern analysis and machine intelligence*, no. 5, pp. 698–700, 1987.

[47] Y. Zhao, X. Zhang, S. Zhang, S. Qiu, H. Yin, and X. Zhang, "Hpcr-vi: Heterogeneous point cloud registration for vehicle-infrastructure collaboration," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–6.

[48] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[49] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun, *et al.*, "V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5486–5495.

[50] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," *International journal on advances in systems and measurements*, vol. 5, no. 3&4, 2012.

[51] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[52] H. Lim, S. Yeon, S. Ryu, Y. Lee, Y. Kim, J. Yun, E. Jung, D. Lee, and H. Myung, "A single correspondence is enough: Robust global registration to avoid degeneracy in urban environments," in *2022 international conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 8010–8017.

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

[54] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[55] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[56] J. Serafin and G. Grisetti, "Using extended measurements and scene merging for efficient and robust point cloud registration," *Robotics and Autonomous Systems*, vol. 92, pp. 91–106, 2017.

[57] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[58] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[59] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

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

[61] 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.