

Robust Collaborative Perception without External Localization and Clock Devices

Zixing Lei¹, Zhenyang Ni¹, Ruize Han², Shuo Tang¹, Chen Feng³, Siheng Chen^{1,4}, Yanfeng Wang^{4,1}

Abstract—A consistent spatial-temporal coordination across multiple agents is fundamental for collaborative perception, which seeks to improve perception abilities through information exchange among agents. To achieve this spatial-temporal alignment, traditional methods depend on external devices to provide localization and clock signals. However, hardware-generated signals could be vulnerable to noise and potentially malicious attack, jeopardizing the precision of spatial-temporal alignment. Rather than relying on external hardware, this work proposes a novel approach: aligning by recognizing the inherent geometric patterns within the perceptual data of various agents. Following this spirit, we propose a robust collaborative perception system that operates independently of external localization and clock devices. The key module of our system, *FreeAlign*, constructs a salient object graph for each agent based on its detected boxes and uses a graph neural network to identify common subgraphs between agents, leading to accurate relative pose and time. We validate *FreeAlign* on both real-world and simulated datasets. The results show that, the *FreeAlign* empowered robust collaborative perception system perform comparably to systems relying on precise localization and clock devices. **Code** will be released.

I. INTRODUCTION

Collaborative perception [1], [2] facilitates the sharing of complementary perceptual information among multiple agents [3], fostering a comprehensive understanding of their surroundings [4]. This approach offers a means to address several inherent limitations of single-agent perception, including occlusion and long-range challenges. Previous methods [5], [6], [7] provides efficacious collaboration techniques and promising performance [8], [9]. Collaborative perception has the potential to be widely used in vehicle-to-everything communication-assisted autonomous driving[10], [11], [12] and multi-UAVs (unmanned aerial vehicles) [13], [14], [15].

A consistent spatial-temporal coordinate system accessible to all agents is the cornerstone of collaborative perception. Traditional methods leverage high-end external devices, like GPS+RTK receivers and synchronized global clocks, to log poses and timestamps, ensuring accurate spatial-temporal alignment. However, this hardware-based manner introduces three significant issues. **Global Localization Noise:** The presence of noise in device outputs (such as GPS) causes inaccurate relative transformation between collaborative messages, leading to performance degradation. While current

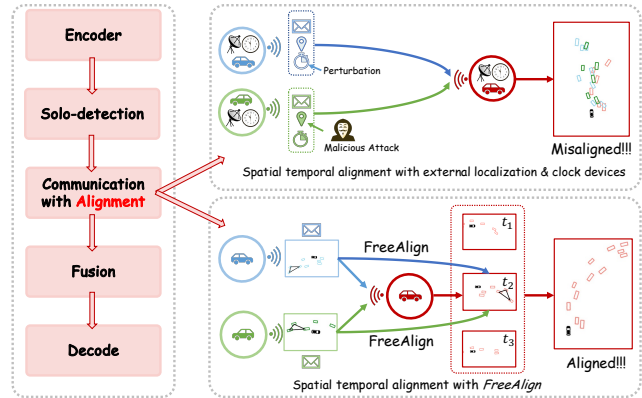


Fig. 1: An illustration of collaborative perception in autonomous driving. The upper side shows in traditional approach, inaccurate localization and clock signal mislead the collaborative perception system. The lower side show the proposed *FreeAlign* enables robust collaborative perception that operates without external devices for localization or synchronized clock. The core of *FreeAlign* is to associate the same objects perceived by multiple agents based on similar geometric structures among those objects.

state-of-the-art practices [16] tried to address this issue, it can only manage minor noise, such as translations less than $1m$, and struggles with more serious noise. **Clock Deviation:** Previous methods such as [4], [17] for collaborative perception often work under the presumption of known latency. Yet, the latency is calculated with timestamps based on clocks of different agents. Therefore, deviation across agents' clocks may produce inaccurate latency measurements, leading to temporal alignment errors. **Vulnerability to Malicious Attacks:** The reliance on external spatial-temporal data, presenting an additional avenue for illusory data injections, makes the system more susceptible to disruptions from malicious attacks, especially through V2V network[18].

In this work, we propose a robust collaborative perception system that operates independently of external devices for localization and clock. Its key module, *FreeAlign*, is a novel spatial-temporal alignment method that leverages graph matching techniques to identify similar geometric patterns within the perceptual data of various agents, ensuring accurate alignment in both spatial and temporal domains; see an illustration in Fig. 1. *FreeAlign* comprises three key components: i) salient-object graph learning, which uses a Graph Neural Network (GNN) to capture comprehensive edge features among the salient objects detected by each agent; ii) multi-anchor-based subgraph searching, which identifies the

¹Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, China. {chezacarss, 0107nzy, tanshu, sihengc, wangyanfeng622}@sjtu.edu.cn.

²Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences rz.han@siat.ac.cn

³New York University, USA. cfeng@nyu.edu

⁴Shanghai AI laboratory, China.

TABLE I: The notations and relation of the important spatial-temporal concepts in collaborative perception.

Notations	Paraphrase
t_i	Current time of ego agent i . Following notations are based on the clock in the agent i .
$t_{j \rightarrow i} / \tilde{t}_{j \rightarrow i}$	Accurate/Estimated(Gotten by external devices) time of collaborative message sent from agent j to ego agent i was perceived.
$\Delta t_{j \rightarrow i} / \tilde{\Delta} t_{j \rightarrow i}$	Accurate/Estimated latency of the collaborative message sent from agent j to agent i , $\Delta t_{j \rightarrow i} = t_i - t_{j \rightarrow i}$, $\tilde{\Delta} t_{j \rightarrow i} = t_i - \tilde{t}_{j \rightarrow i}$.
$\delta t_{j \rightarrow i}$	Clock deviation between agent i and agent j . $\delta t_{j \rightarrow i} = t_{j \rightarrow i} - \tilde{t}_{j \rightarrow i} = \Delta t_{j \rightarrow i} - \tilde{\Delta} t_{j \rightarrow i}$. δt_j will harm the compensation module.
τ	Sample interval in collaborative perception system ($\leq 100\text{ms}$)
$\xi_i^t / \tilde{\xi}_i^t$	Accurate/Estimated(Gotten by external devices) pose of agent i at time t_i in the global coordinate system.
$\Delta \xi_{j \rightarrow i}^t / \tilde{\Delta} \xi_{j \rightarrow i}^t$	Accurate/Estimated relative pose between agent j at time $t_{j \rightarrow i}$ and agent i at time t_i , $\Delta \xi_{j \rightarrow i}^t = f_{\text{trans}}(\xi_i^t, \xi_j^{t_{j \rightarrow i}})$.
$\delta \xi_{j \rightarrow i}^t$	Relative pose error between agent j at time $t_{j \rightarrow i}$ and agent i at time t_i , $\delta \xi_{j \rightarrow i}^t = \Delta \xi_{j \rightarrow i}^t - \tilde{\Delta} \xi_{j \rightarrow i}^t = (\delta x_{j \rightarrow i}^t, \delta y_{j \rightarrow i}^t, \delta \theta_{j \rightarrow i}^t)$.
\mathcal{O}_i	Odometry of agent i . Used to calculate the relative pose among agent i 's observations at different times.

proximate maximum common subgraph across two salient-object graphs; iii) relative transformation calculation, which leverages the common subgraph to calculate the relative pose and latency between two collaborative messages.

The advantages of our system are twofold: i) it provides a machine learning approach to substitute global localization and synchronized devices, substantially bolstering the robustness of collaborative perception; and ii) the key component *FreeAlign*, allows for seamless integration with numerous established methods to compose our system without necessitating retraining of the collaborative perception architecture.

In summary, the main contribution of this work are:

- We propose the first collaborative perception system without relying on external localization and clock devices;
- We propose *FreeAlign*, a novel graph-matching-based spatial-temporal alignment method that can be seamlessly integrated into existing collaborative perception systems.
- We conduct extensive experiments for collaborative LiDAR-based object detection in simulated and real-world datasets to validate that, integrating *FreeAlign* leads to considerable improvements for a majority of previous methods when facing the typical issues.

II. RELATED WORK

Collaborative Perception. The rise of multi-agent collaborative perception addresses limitations inherent to single-agent perception, such as occlusion. Several datasets, including V2X-Sim[2], OPV2V[3], and DAIR-V2X[4], contribute to the body of research. Leveraging these datasets, a host of methods have emerged. For instance, V2VNet[1] integrates collaboration into perception and prediction. Studies such as Who2comm [19], When2comm [20], and Where2comm[11] optimize perception performance against bandwidth efficiency, while SyncNet [17] addresses communication latency. However, these approaches often assume reliable input from localization and clock devices, assuming negligible noise interference. In contrast, our work envisions collaborative perception independent of such external devices.

Graph Matching. Graph matching and subgraph search processes are pivotal in various domains, and this study uses them for spatial-temporal alignment. Algorithms like VF2[21], QuickSI[22], VF2x[23], VF2 Plus[24], and VF3[25] employ feasibility functions and depth-first search for graph isomorphism challenges, although they have limitations with graphs of varying node numbers. SuperGlue[26] utilizes a neural framework to match local features through a differentiable optimal transport solution. PCA-GM[27]

focuses on node embeddings and affinity metrics, but can overfit under certain conditions. These existing methods often falter in collaborative perception due to an over-reliance on visual features and pixel coordination. In contrast, our proposed approach tailors specifically to collaborative perception, emphasizing invariant geometric interrelations. For effective feature extraction, we adopt graph representation learning, incorporating techniques like node2vec[28], deepwalk[29], and GNN methods such as GCN[30], GAT[31], GATv2[32], and EGAT[33]. Particularly, our implementation utilizes EGAT, which accentuates edge features for enhanced subgraph search reliability.

III. COLLABORATIVE PERCEPTION SYSTEM

Consider N agents collaboratively detecting 3D objects within the scene. Let \mathbf{X}_i^t be the observation of the i -th agent at time t , \mathbf{B}_i^t and $\hat{\mathbf{B}}_i^t$ be the detected boxes before and after collaboration, respectively. For the i -th agent, the proposed collaborative 3D object detection works as follows:

$$\mathbf{F}_i^{t_i}, \mathbf{B}_i^{t_i} = f_{\text{solo-detection}}(\mathbf{X}_i^{t_i}), \quad (1a)$$

$$\tilde{\Delta} \xi_{j \rightarrow i}^t, \tilde{\Delta} t_{j \rightarrow i} = f_{\text{FreeAlign}}\left(\left\{\mathbf{B}_i^{t_i - k\tau}\right\}_{k=0,1,\dots}, \mathbf{B}_j^{t_{j \rightarrow i}}, \mathcal{O}_i\right), \quad (1b)$$

$$\tilde{\mathbf{F}}_{j \rightarrow i}^{t_{j \rightarrow i}}, \tilde{\mathbf{B}}_{j \rightarrow i}^{t_{j \rightarrow i}} = f_{\text{spatial-trans}}\left(\tilde{\Delta} \xi_{j \rightarrow i}^t, \mathbf{F}_j^{t_{j \rightarrow i}}, \mathbf{B}_j^{t_{j \rightarrow i}}\right), \quad (1c)$$

$$\hat{\mathbf{F}}_{j \rightarrow i}^{t_i} = f_{\text{temporal-trans}}\left(\left\{\tilde{\Delta} t_{j \rightarrow i}, \tilde{\mathbf{F}}_{j \rightarrow i}^{t_{j \rightarrow i} - k\tau}\right\}_{k=0,1,\dots}\right), \quad (1d)$$

$$\hat{\mathbf{F}}_i^{t_i} = f_{\text{fusion}}\left(\mathbf{F}_i^{t_i}, \left\{\hat{\mathbf{F}}_{j \rightarrow i}^{t_i}\right\}_{j \in \mathcal{N}_i}\right), \quad (1e)$$

$$\hat{\mathbf{B}}_i^{t_i} = f_{\text{decoder}}\left(\hat{\mathbf{F}}_i^{t_i}\right), \quad (1f)$$

where \mathbf{F}_i^t and $\hat{\mathbf{F}}_i^t$ are the i -th agent's feature maps before and after collaboration at time t , respectively, $\tilde{\mathbf{F}}_{j \rightarrow i}^t$ is the feature map transmitted from Agent j to i at time t , $\tilde{\Delta} \xi_{j \rightarrow i}^t$ and $\tilde{\Delta} t_{j \rightarrow i}$ are the relative pose and time between Agents i and j , and \mathcal{N}_i is the collaborator of the i -th agent.

In Step (1a), Agent i encodes its observed data $\mathbf{X}_i^{t_i}$ into a BEV feature map $\mathbf{F}_i^{t_i}$, thereafter producing detected boxes $\mathbf{B}_i^{t_i}$. The agent subsequently exchanges this information with its collaborators \mathcal{N}_i . Notably, the exact time $t_{j \rightarrow i}$ remains **unknown** to agent i , while the estimated time $t_{j \rightarrow i}$ is noisy due to clock deviations. In Step (1b), Agent i computes the relative pose and latency between its historical observations and each of its collaborators by utilizing proposed *FreeAlign*, achieving spatial-temporal alignment. In Step (1c), Agent i transforms the $\mathbf{F}_j^{t_{j \rightarrow i}}, \mathbf{B}_j^{t_{j \rightarrow i}}$ to its own spatial coordinates at time t_i , resulting in $\tilde{\mathbf{F}}_{j \rightarrow i}^{t_{j \rightarrow i}}, \tilde{\mathbf{B}}_{j \rightarrow i}^{t_{j \rightarrow i}}$. Note that $\tilde{\mathbf{F}}_{j \rightarrow i}^{t_{j \rightarrow i}}, \tilde{\mathbf{B}}_{j \rightarrow i}^{t_{j \rightarrow i}}$ are noisy since $\tilde{\Delta} \xi_{j \rightarrow i}^t$ is estimated. In Step (1d), Agent i com-

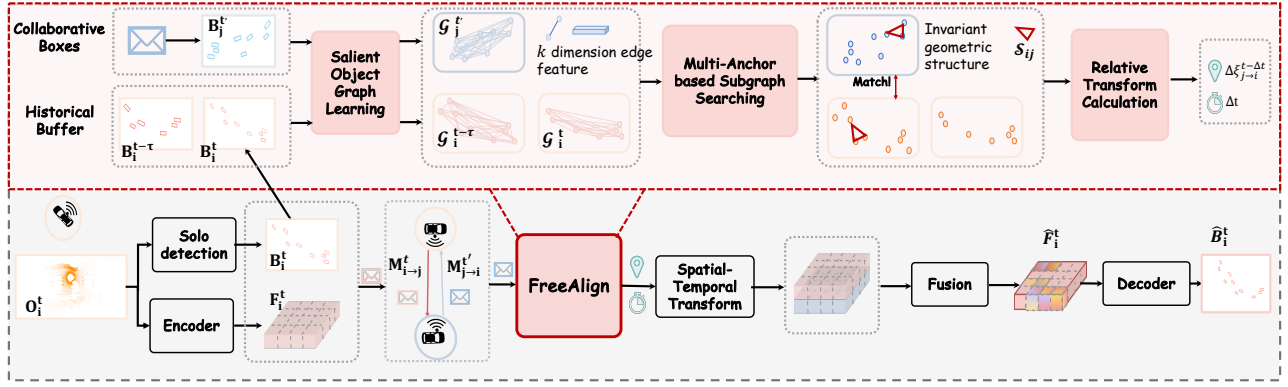


Fig. 2: Overview of the proposed robust collaborative perception framework. The key module is *FreeAlign*, which leverages salient-object graphs to achieve spatial-temporal alignment.

pensates for potential noisy latencies using past collaborative feature maps. In Step (1e), Agent i aggregates all compensated feature maps to yield a consolidated feature map $\hat{\mathbf{F}}_i^{t_i}$, which generates the final detections $\hat{\mathbf{B}}_i^{t_i}$ in Step (1f).

Precise spatial-temporal alignment is the foundation of effective collaborative perception. A misalignment in this step would significantly damage the subsequent procedures, including feature transformation, fusion and collaborative detection. Traditionally, to achieve the spatial-temporal alignment as Step. (1b), each agent needs an external localization and clock devices to receive signals and provide its globally synchronized pose and time, subsequently computing relative transformations between collaborative messages. Due to the dependence of noisy external signals, this hardware-based approach could cause numerous issues, such as localization noise, clock deviation and vulnerability to malicious attacks. Our proposed method, *FreeAlign*, specifically addresses this by emphasizing perceptual data, utilizing graph matching and machine learning for alignment. A detailed discussion on this will be presented in the following section.

IV. FREEALIGN: GRAPH MATCHING BASED SPATIAL-TEMPORAL ALIGNMENT

To reform the traditional approach, our intuition is that the geometric relation of shared objects (in the real world) between two collaborative messages is unified if both messages are perceived simultaneously, thereby allowing spatial alignment according to the geometric relation and latency determination through a time-series search. Motivated by this, we present *FreeAlign*, a graph matching based spatial-temporal alignment approach. *FreeAlign* adeptly discerns the consistent geometric structure across collaborative messages, facilitating accurate computation of relative transformation in the spatial-temporal continuum. This method comprises three essential modules, i.e., graph learning, subgraph searching, and transformation calculation, which are presented as below.

A. Salient-Object Graph Learning

To identify similar geometric structures between two collaboration messages, a metric for assessing similarity is required. An intuitive approach is to use the attributes of

the objects detected by agents. Being invariant to coordinate systems and more distinctive, relative distance serves as a superior metric. However, sole reliance on it can result in confusion due to noise and coincidental matches. To refine the attribute for matching, we propose to model all boxes perceived by one agent as a salient-object graph, where each node is one box. A GNN can then be employed to learn edge features through message passing, enhancing their comprehensiveness and expressiveness by incorporating information from both relative distance and geometric structure. This, in turn, facilitates a more effective identification process.

For the i -th agent at timestamp t , let $\mathcal{G}_i^t = (\mathcal{V}_i^t, \mathcal{E}_i^t)$ be a fully-connected salient-object graph, where \mathcal{V}_i^t is the set of n nodes with each node one salient object detected by agent i and \mathcal{E}_i^t is the edge set modeling all the pairwise relationships between salient objects. Here we aim to learn a tensor of edge features $\mathbf{W}_i^t \in \mathbb{R}^{n \times n \times k}$, where each pillar is the edge feature $(\mathbf{W}_i^t)_{pq} \in \mathbb{R}^k$ that reflects the invariant proximity between two salient objects p, q . Here the invariance means that whatever an agent's pose is, the edge feature of two salient objects is unaltered. This implies that when both agents i and j detect the same two salient objects at the same time t , the corresponding edge features in \mathbf{W}_i^t and \mathbf{W}_j^t would be the same. Let $\mathbf{R}_i^t \in \mathbb{R}^{n \times n}$ is a matrix that encodes the relative distance between each pair of nodes, to learn such an edge feature, we use EdgeGAT [33] as the graph representation learning model, that is, $\mathbf{W}_i^t = f_{\text{EdgeGAT}}(\mathcal{G}_i^t, \mathbf{R}_i^t)$.

For the training of the EdgeGAT model, we employ a contrastive learning loss which can be formulated as follows

$$L = \frac{1}{|\mathcal{M}|} \sum_{(p,q) \sim \mathcal{M}} \left\| (\mathbf{W}_i^t)_{pq} - (\mathbf{W}_j^{t'})_{ef} \right\|_2 + \frac{1}{|\mathcal{U}|} \sum_{(u,v) \sim \mathcal{U}} \max \left(\gamma - \left\| (\mathbf{W}_i^t)_{pq} - (\mathbf{W}_j^{t'})_{uv} \right\|_2, 0 \right), \quad (2)$$

where $(\mathbf{W}_i^t)_{pq}$ and $(\mathbf{W}_j^{t'})_{ef}$ represent the edge feature between nodes p and q in \mathcal{G}_i^t , and e and f in $\mathcal{G}_j^{t'}$, respectively. (p, q) denotes an edge in the matched edge set \mathcal{M} and edge (p, q) matches edge (e, f) . (u, v) denotes an edge in the unmatched edge set \mathcal{U} , which is sampled from all edges

without match. γ serves as the margin. This contrastive loss aids the GNN in learning distinctive edge features that facilitate the identification of matches between two graphs.

B. Multi-Anchor Based Subgraph Searching

Upon completion of the salient-object graph learning phase, the task is to locate the approximate maximum common subgraph between two salient-object graphs, signifying distinct and similar geometric structures. Let some potentially matched nodes pairs be anchors. The core concept is that nodes are deemed a match if all edges linking the node and anchors in both graphs are similar, thereby forming a stable common subgraph with all anchors within its graph. Let \mathcal{S} be the common subgraph and ε be the confident score for this common subgraph, finding common subgraph between two salient-object graph $\mathcal{G}_i, \mathcal{G}_j$ whose edge feature tensors are $\mathbf{W}_i, \mathbf{W}_j$ by MASS can be formulated as

$$\mathcal{S}, \varepsilon = f_{\text{MASS}}(\mathcal{G}_i, \mathbf{W}_i, \mathcal{G}_j, \mathbf{W}_j). \quad (3)$$

When there are n and m nodes in the \mathcal{G}_i and \mathcal{G}_j respectively. In order to realize $f_{\text{MASS}}(\cdot)$, the procedure is divided into four primary steps:

i). Initialization. We begin by generating potential anchor lists. This involves $n \times m$ pairs of potential matching nodes to serve as the initial anchor for subsequent searches.

ii). Anchor lists expanding. For each initialized matched pair (p, q) among the $n \times m$ potential node pairs, there is only one pair of anchors, which easily leads to instability given the various situations. To mitigate this, we expand the anchor lists. By examining another node pair (u, v) and comparing the edge feature $\mathbf{W}_i(p, u)$ and $\mathbf{W}_j(q, v)$, we can ascertain the similarity. If the difference, denoted as $\varepsilon_{(p,u),(q,v)} = |\mathbf{W}_i(p, u) - \mathbf{W}_j(q, v)|$ is below a predefined threshold, then the pair (u, v) will be added into the anchor list. is incorporated into the anchor list. This expansion proceeds until the count of anchor pairs meets the limit γ ensuring that edge features related to all anchors are assessed.

iii). Subgraph search. For every potential anchor list, nodes are incrementally added to the subgraph \mathcal{S} until no other nodes meet the inclusion criteria.

iv). Selection. The discrepancy value, ε , is computed as $\varepsilon = \frac{1}{r^p} \sum \varepsilon_e$ where r represents the size of subgraph \mathcal{S} and p acts as a tunable hyperparameter. Out of all the subgraphs, the one with the minimal ε is chosen to represent the common subgraph.

C. Relative Transformation Calculation

With MASS, we can get a common subgraph \mathcal{S} and its score ε between two salient-object graphs. To obtain the relative pose and time difference in (1b), we need to calculate the relative transformation according to the common subgraph. It takes two step in temporal and spatial domain, respectively.

Clock deviation estimation To determine Δt , we leverage MASS between $\mathcal{G}_j^{t_j \rightarrow i}$ and a temporal salient-object graph buffer $[\mathcal{G}_i^{t_i}, \mathcal{G}_i^{t_i - \tau}, \dots, \mathcal{G}_i^{t_i - l\tau}]$ in the i -th agent. After obtaining a list of matching results at different timestamps, the most reliable match is selected via the ε . If a collaborative message

fails to identify a common subgraph, whose number of nodes should exceed a predetermined minimum threshold, across the time buffer, *FreeAlign* will discard this collaborative perception message to ensure safety.

Relative pose estimation. To calculate $\Delta \xi_{j \rightarrow i}^t$, the common subgraph selected in temporal alignment is utilized to derive the relative pose $(\Delta x, \Delta y, \Delta \theta)$ as the common subgraph \mathcal{S}_{ij} can be considered as two sets of matching points $\mathbf{S}_i^t, \mathbf{S}_j^t \in \mathbb{R}^{\psi \times 3}$ in two coordinate systems with robust RANSAC[35] or LMesS[36] algorithm.

D. Discussions

Advantages. Standard general alignment methods, such as SuperGlue[26] and PCA-GM[27] cannot be applied to collaborative perception for two reasons: i) they leverage a keypoint detection network to extract key points as nodes with visual features, which are not provided in collaborative perception; ii) they leverage pixel coordinates to calculate edge attributes, which is highly affected by rotations. In comparison, *FreeAlign* i) focus on geometric relationships of objects; and ii) leverages the relative distances, which are invariant from different perspectives.

Prerequisites. There are two assumptions for *FreeAlign* to function well. i) Collaboration is initiated only when agents are in close proximity, ensuring a common field of view. ii) The scenario is dynamic. This makes it almost impossible for the geometric patterns at varying timestamps to be identical, allowing us to determine the timestamp. *These assumptions are typically common and practical in a collaborative perception setting.*

FreeAlign and GNSS Integration: The *FreeAlign* operates independently of GNSS inputs but can benefits from their integration. Firstly, it assesses the fundamental reliability of GNSS signal. Subsequently, GNSS enhances *FreeAlign*'s efficiency by reducing its search range.

V. EXPERIMENTAL RESULTS

A. Datasets and Experimental Settings

We conduct collaborative LiDAR-based 3D object detection on both a simulation dataset, OPV2V [3], co-simulated by OpenCDA [5] and Carla [37], and a real-world dataset, DAIR-V2X [4]. The detection results are evaluated by Average Precision (AP) at Intersection-over-Union (IoU) threshold of 0.30 and 0.50. We follow [38] and [16] to set the detection range as $x \in [-140m, 140m], y \in [-40m, 40m]$ in OPV2V and $x \in [-100m, 100m], y \in [-40m, 40m]$ in DAIR-V2X, respectively. We use PointPillars [39] with the grid size $(0.4m, 0.4m)$ as the encoder. For multi-scale feature fusion, the residual layer number is 3 and the channel numbers are $(64, 128, 256)$.

B. Quantitative Evaluation

Detection performance in presence of pose errors. Fig. 3 compares detection performances of *FreeAlign*'s with the previous methods under varying pose noise levels on both OPV2V and DAIR-V2X datasets. The baselines include single-agent detection, V2VNet[1], DiscoNet[10],

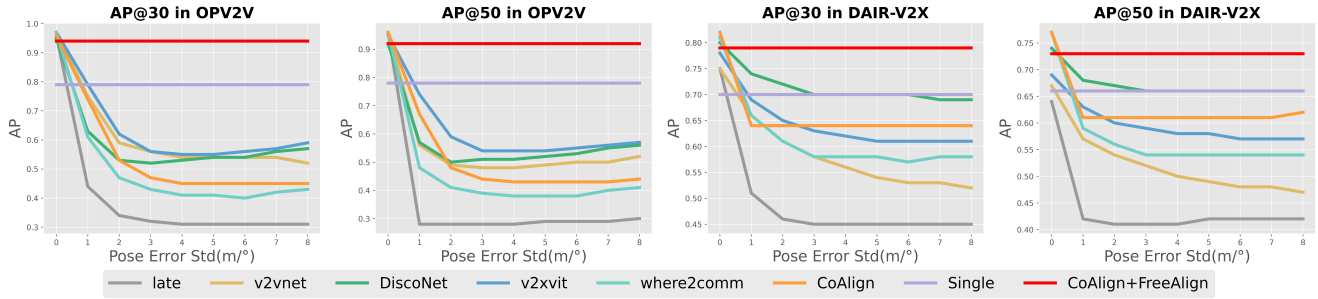


Fig. 3: Detection performance on OPV2V [3] and DAIR-V2X[4] datasets with pose noises following Gaussian distribution in the testing phases. The performance of baselines drops significantly as noise increases, while *FreeAlign*'s result is stable.

TABLE II: Comparison between *FreeAlign* and other matching methods. Only *FreeAlign* can achieve accurate matching in the collaborative perception scenarios.

	OPV2V			DAIR-V2X		
	$\delta\theta(^{\circ})$	$\sqrt{(\delta x)^2 + (\delta y)^2}(\text{m})$	Error rate	$\delta\theta(^{\circ})$	$\sqrt{(\delta x)^2 + (\delta y)^2}(\text{m})$	Error rate
ICP[34]	61.91	24.75	93.90%	79.82	43.25	96.32%
SuperGlue[26]	52.05	22.75	90.94%	57.04	80.44	96.61%
PCA-GM[27]	41.48	13.65	90.27%	55.81	50.70	99.43%
FreeAlign	0.017	0.266	0.56%	0.015	0.318	0.46%

TABLE III: Collaborative detection under various latency deviations. *FreeAlign* is robust to latency deviations.

$\delta t_{j \rightarrow i}$ (clock deviation)	0	100ms	200ms	300ms
Single	0.78	0.78	0.78	0.78
Where2comm	0.66	0.66	0.66	0.66
Where2comm+SyncNet	0.87	0.86	0.68	0.57
Ours	0.86	0.86	0.86	0.86

TABLE IV: *FreeAlign* deliver accurate temporal alignment in OPV2V and DAIR-V2X datasets.

	Synchronization Accuracy	Average $\delta t_{j \rightarrow i}$
OPV2V	79.46%	22.8ms
DAIR-V2X	66.36%	45.6ms

V2X-ViT[6], Where2comm[11], CoAlign[16] and late fusion. Gaussian noise $\mathcal{N}(0, \sigma)$ is applied to (x, y, θ) , with $\sigma \in [0, 8]$. We can see that *FreeAlign*'s performance is robust, almost unaffected by pose noises, while the baseline methods show significant performance drops with pose noises. *FreeAlign* improves detection by 20.5% compared to single-agent perception under serious pose noises.

Detection performance in the presence of latency deviation. To test *FreeAlign* in the presence of latency deviation (only know a noisy latency value), Table III displays the AP 0.5 metric for the OPV2V dataset under 400ms latency with a noisy estimation, which means the compensation model only compensate for an inaccurate latency. The results show that i) without latency compensation, Where2comm's performance falls below that of a single agent; ii) SyncNet can compensate given an accurate latency value, but experiences a performance decline with increasing deviation; iii) our methods, Where2comm+SyncNet+FreeAlign can estimate a more accurate latency value, resulting in a notable gain over single-agent detection. The performances of single detection and Where2comm remain constant regardless of latency values as they do not compensate for latency. With the same setting, Table IV further exhibits that *FreeAlign* achieves an accurate temporal alignment in both datasets. Note that

TABLE V: With *FreeAlign*, all collaboration methods significantly improve their robustness to localization attack.

	OPV2V		DAIR-V2X	
	AP@0.3	AP@0.5	AP@0.3	AP@0.5
Single	0.79	0.78	0.70	0.66
Late	0.56	0.54	0.40	0.38
V2VNet[1] ECCV2020	0.44	0.40	0.42	0.39
DiscoNet[10] NeurIPS2021	0.56	0.47	0.64	0.54
V2X-ViT[6] ECCV2022	0.60	0.59	0.55	0.52
Where2comm[11] NeurIPS2022	0.32	0.30	0.46	0.44
CoAlign[16] ICRA2023	0.32	0.30	0.64	0.61
V2VNet+FreeAlign	0.88 \uparrow 0.44	0.75 \uparrow 0.35	0.69 \uparrow 0.27	0.60 \uparrow 0.21
DiscoNet+FreeAlign	0.87 \uparrow 0.31	0.79 \uparrow 0.32	0.76 \uparrow 0.12	0.70 \uparrow 0.16
V2X-ViT+FreeAlign	0.93 \uparrow 0.33	0.89 \uparrow 0.30	0.73 \uparrow 0.18	0.65 \uparrow 0.13
Where2comm+FreeAlign	0.86 \uparrow 0.54	0.71 \uparrow 0.41	0.75 \uparrow 0.29	0.67 \uparrow 0.23
CoAlign+FreeAlign	0.95\uparrow 0.63	0.94\uparrow 0.64	0.79\uparrow 0.15	0.73\uparrow 0.12

- i) Synchronization Accurate means the *FreeAlign* find the correct timestamp and see Table I for the meaning of $\delta t_{j \rightarrow i}$.
- ii) there is no latency compensation module in Where2comm, so it will not be affected by clock deviation.

Detection performance under localization attack. We also explore detection performance under a malicious attack, a scenario where deliberately misleading pose data aims to generate false positive bounding boxes. Table V presents the AP 0.3 and AP 0.5 of various methods under this attack. The results reveal that *FreeAlign* significantly improves performance, enhancing CoAlign by 206.7% on the OPV2V dataset, for instance. Notably, with *FreeAlign*'s assistance, a majority of collaborative perception methods outperform single-agent perception even under malicious attack.

Matching performance. To validate the performance on subgraph matching, we compare *FreeAlign* with ICP[34], SuperGlue[26], and PCA-GM[27]. Table II shows the error between the estimated pose and the ground truth. Error rate means the proportion of cases that have transformation error greater than 3m (which makes an accurately detected box be regarded as false positive even in AP 0.3 metric). We see that the other methods fail to match the corresponding

TABLE VI: Ablation studies on OPV2V and DAIR-V2X datasets.

Dataset		OPV2V			DAIR-V2X		
Anchor Based	GNN Feature	$\delta\theta(^{\circ})$	$\sqrt{(\delta x)^2 + (\delta y)^2}(\text{m})$	Error Rate	$\delta\theta(^{\circ})$	$\sqrt{(\delta x)^2 + (\delta y)^2}(\text{m})$	Error Rate
		4.657	0.979	3.71%	15.356	13.887	31.39%
	✓	2.186	0.904	2.46%	5.721	4.962	9.59%
✓		0.029	0.283	0.78%	0.032	0.636	1.38%
✓	✓	0.017	0.266	0.56%	0.015	0.318	0.46%

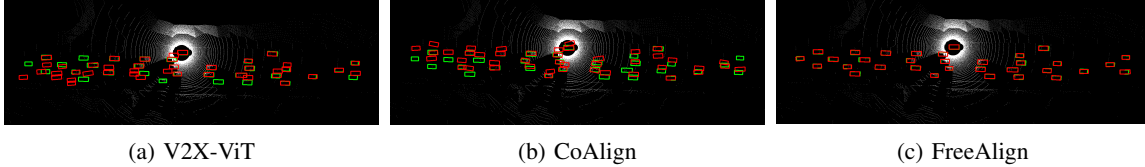


Fig. 4: *FreeAlign* qualitatively outperforms V2X-ViT and CoAlign on OPV2V dataset under pose noisy setting. Green and red boxes denote ground-truth and detection, respectively.

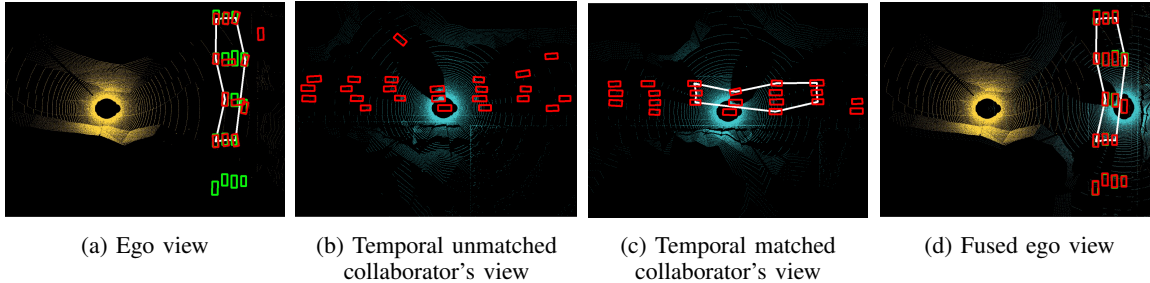


Fig. 5: Visualization of collaboration between ego vehicle (a) and edge vehicle (b-c), and the spatial-temporal fusion result of them (d). *FreeAlign* achieves spatial-temporal alignment through a common subgraph between views of different times and location. Green and red boxes denote ground-truth and detection, respectively. Yellow and blue denote the point clouds collected from ego vehicle and edge vehicle, respectively. White denotes the common subgraph.

subgraph and only the proposed *FreeAlign* can achieve accurate matching in the collaborative perception scenarios.

C. Ablation Studies and Discussions

Table VI assesses the effectiveness of the proposed anchor-based matching and edge features learned by GNN. Absent the GNN features, edge matching is determined by the relative distance between two nodes. Without employing the anchor-based method, the search is anchored to a single pair. More results can be found in the supplementary material. We find that 1) the anchor-based matching method is critical to our matching method. In DAIR-V2X dataset, without the multi-anchor setting, the error rate increases nearly 20 times; and 2) edge features learned by GNN help avoid failure cases, especially in a noisy real-world dataset DAIR-V2X. Results show the proposed anchor-based matching and GNN-based edge features make *FreeAlign* more robust to noisy real-world applications.

D. Qualitative Evaluations

Visualization of detection results. Fig. 4 shows a comparative visualization of detection results from V2X-ViT, CoAlign, and *FreeAlign* in the OPV2V dataset under noisy setting. The noise stems from a Gaussian distribution with a standard deviation of 3.0m for position and 3.0° for heading. V2X-ViT, despite employing the MSWin module to mitigate pose error, struggles under large noise. Similarly, the pose graph optimization algorithm of CoAlign fails in the presence of large noise, leading to a significant drop in detection

performance. In contrast, *FreeAlign*'s exhibits superior performance under large noise. This can be attributed to its independence from prior pose information, which makes it less susceptible to the impacts of pose noise.

Visualization of spatial-temporal alignment. Fig. 5 offers a comprehensive visual depiction of *FreeAlign*'s mechanism in a specific scenario, where the ego vehicle is located at a T-junction, limiting visibility. Fig. 5a displays the ego vehicle's perspective, illustrating suboptimal detection performance. Fig. 5b displays the collaborator's perspective at an unsynchronized timestep, rendering it incompatible with the observation in Fig. 5a. Fig. 5c displays the collaborator's view at a synchronized timestep with ego. Note that there exist invariant geometric structures between Figs. 5a and 5c. Fig. 5d demonstrates that, with the *FreeAlign*'s assistance, the ego vehicle successfully detects through the T-junction.

VI. CONCLUSION

This paper proposed *FreeAlign*, a novel spatial-temporal alignment method for robust collaborative perception without any expensive high-end global localization and synchronized clock. The core idea is to leverage the invariant geometric structure composed of objects commonly perceived and observed by the individual agents to estimate the relative poses and time differences. Comprehensive experiments covering simulated and real-world datasets showed that *FreeAlign* achieve efficient and accurate spatial-temporal alignment and help collaborative perception to be significantly more robust.

REFERENCES

- [1] 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.
- [2] 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. 10914–10921, 2022.
- [3] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.
- [4] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, et al., "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.
- [5] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "Openeda: an open cooperative driving automation framework integrated with co-simulation," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 1155–1162.
- [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 *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*. Springer, 2022, pp. 107–124.
- [7] 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," *arXiv preprint arXiv:2305.05938*, 2023.
- [8] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.
- [9] 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.
- [10] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.
- [11] 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*.
- [12] S. Chen, B. Liu, C. Feng, C. Vallespi-Gonzalez, and C. Wellington, "3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 68–86, 2020.
- [13] Y. Hu, S. Fang, W. Xie, and S. Chen, "Aerial monocular 3d object detection," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 1959–1966, 2023.
- [14] E. T. Alotaibi, S. S. Alqefari, and A. Koubaa, "Lsar: Multi-uav collaboration for search and rescue missions," *IEEE Access*, vol. 7, pp. 55 817–55 832, 2019.
- [15] A. Alagha, S. Singh, R. Mizouni, J. Bentahar, and H. Otrok, "Target localization using multi-agent deep reinforcement learning with proximal policy optimization," *Future Generation Computer Systems*, vol. 136, pp. 342–357, 2022.
- [16] Y. Lu, Q. Li, B. Liu, M. Dianat, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," *arXiv preprint arXiv:2211.07214*, 2022.
- [17] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 2022, pp. 316–332.
- [18] B. K. Bhargava, A. M. Johnson, G. I. Munyengabe, and P. Angin, "A systematic approach for attack analysis and mitigation in v2v networks," *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, vol. 7, no. 1, pp. 79–96, 2016.
- [19] 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.
- [20] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.
- [21] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub) graph isomorphism algorithm for matching large graphs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 10, pp. 1367–1372, 2004.
- [22] H. Shang, Y. Zhang, X. Lin, and J. X. Yu, "Taming verification hardness: an efficient algorithm for testing subgraph isomorphism," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 364–375, 2008.
- [23] Q. Yin and T. Roscoe, "Vf2x: Fast, efficient virtual network mapping for real testbed workloads," in *Testbeds and Research Infrastructure. Development of Networks and Communities*, T. Korakis, M. Zink, and M. Ott, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 271–286.
- [24] V. Carletti, P. Foggia, and M. Vento, "Vf2 plus: An improved version of vf2 for biological graphs," in *Graph-Based Representations in Pattern Recognition*, C.-L. Liu, B. Luo, W. G. Kropatsch, and J. Cheng, Eds. Cham: Springer International Publishing, 2015, pp. 168–177.
- [25] V. Carletti, P. Foggia, A. Saggese, and M. Vento, "Introducing vf3: A new algorithm for subgraph isomorphism," in *Graph-Based Representations in Pattern Recognition*, P. Foggia, C.-L. Liu, and M. Vento, Eds. Cham: Springer International Publishing, 2017, pp. 128–139.
- [26] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.
- [27] R. Wang, J. Yan, and X. Yang, "Learning combinatorial embedding networks for deep graph matching," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3056–3065.
- [28] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
- [29] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.
- [30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2019.
- [31] P. Velićković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks."
- [32] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2021.
- [33] K. Kamiński, J. Ludwiczak, M. Jasiński, A. Bukala, R. Madaj, K. Szczepaniak, and S. Dunin-Horkawicz, "Rossmann-toolbox: a deep learning-based protocol for the prediction and design of cofactor specificity in rossmann fold proteins," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab371, 2022.
- [34] 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.
- [35] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981. [Online]. Available: <https://doi.org/10.1145/358669.358692>
- [36] J. M. Steele and W. L. Steiger, "Algorithms and complexity for least median of squares regression," *Discret. Appl. Math.*, vol. 14, no. 1, pp. 93–100, 1986. [Online]. Available: [https://doi.org/10.1016/0166-218X\(86\)90009-0](https://doi.org/10.1016/0166-218X(86)90009-0)
- [37] 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.
- [38] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.
- [39] 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.