# V2VLoc: Robust GNSS-Free Collaborative Perception via LiDAR Localization

**Wenkai Lin**[1,2*], **Qiming Xia**[1,2*], **Wen Li**[1,2], **Xun Huang**[1,2], **Chenglu Wen**[1,2†]

[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing

## Abstract

Multi-agents rely on accurate poses to share and align observations, enabling a collaborative perception of the environment. However, traditional GNSS-based localization often fails in GNSS-denied environments, making consistent feature alignment difficult in collaboration. To tackle this challenge, we propose a robust GNSS-free collaborative perception framework based on LiDAR localization. Specifically, we propose a lightweight Pose Generator with Confidence (PGC) to estimate compact pose and confidence representations. To alleviate the effects of localization errors, we further develop the Pose-Aware Spatio-Temporal Alignment Transformer (PASTAT), which performs confidence-aware spatial alignment while capturing essential temporal context. Additionally, we present a new simulation dataset, V2VLoc, which can be adapted for both LiDAR localization and collaborative detection tasks. V2VLoc comprises three subsets: Town1Loc, Town4Loc, and V2VDet. Town1Loc and Town4Loc offer multi-traversal sequences for training in localization tasks, whereas V2VDet is specifically intended for the collaborative detection task. Extensive experiments conducted on the V2VLoc dataset demonstrate that our approach achieves state-of-the-art performance under GNSS-denied conditions. We further conduct extended experiments on the real-world V2V4Real dataset to validate the effectiveness and generalizability of PASTAT.

**Code** — https://github.com/wklin214-glitch/V2VLoc
**Datasets** — https://huggingface.co/datasets/linwk/V2VLoc

## Introduction

Collaborative perception is a paradigm in which multi-agents share information and cooperate on perception tasks to achieve improved accuracy, wider coverage. Recently, collaborative perception has experienced rapid development (Hu et al. 2022, 2024; Song et al. 2025) and has demonstrated promising performance.

The primary capability of collaborative perception lies in its ability to integrate observations from various viewpoints into a unified coordinate frame. This integration process relies on precise pose estimation. Recent studies (Xu et al. 2022b; Lu et al. 2023; Zhang et al. 2024) have attempted to improve robustness by mitigating the effects of localization errors. However, traditional GNSS-based localization methods, such as GPS combined with RTK receivers, are inherently device-dependent and are prone to degraded or lost signals in GNSS-denied environments (e.g., tunnels, spoofing, jamming, or satellite occlusion). This dependency presents significant challenges for effective collaborative perception.

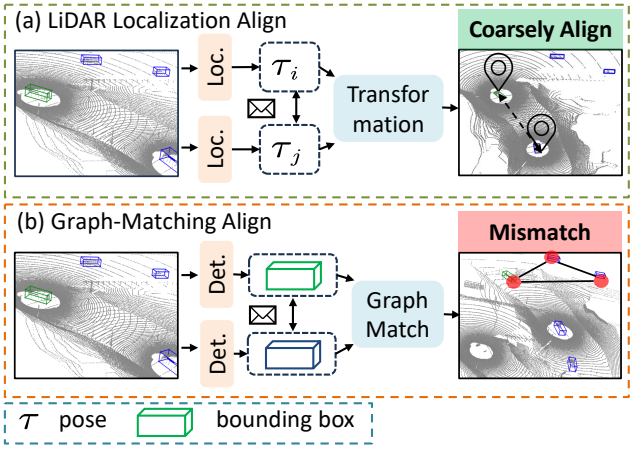


Figure 1: An illustration of different alignment methods. (a) shows that LiDAR localization achieves coarse alignment successfully. (b) shows that the Graph-Matching method fails without consensus objects.

To tackle this issue, FreeAlign (Lei et al. 2024) proposed an alignment method that utilizes the intrinsic geometric patterns present in the sensor data for inter-agent alignment and relative pose estimation. However, as illustrated in Fig. 1(b), this graph-matching-based method relies on bounding box sharing, pairwise greedy graph matching, and assumes the presence of a certain number of co-viewed objects, which makes it less effective in scenarios with sparse or minimally overlapping observations, ultimately leading to unstable pose alignment.

Recent advances have been made in LiDAR-based visual localization models (Choy, Park, and Koltun 2019; Xia et al. 2021; Zhang et al. 2023; Li et al. 2023; Yang et al. 2024a; Li et al. 2024a; Yu et al. 2023b; Li et al. 2025),

*These authors contributed equally.
†Corresponding author.

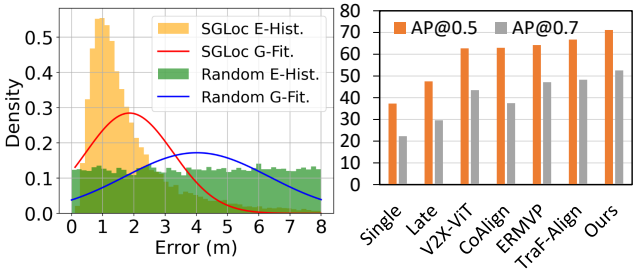


Figure 2: The left figure shows the error distribution of SGLoc (Li et al. 2023) and random noise; the right shows detection performance under LiDAR Localization-induced pose errors. E-Hist: Error Histogram, G-Fit: Gaussian Fit.

which estimate global coordinates using LiDAR data. Compared to map-based methods with heavy storage and communication costs, regression-based LiDAR localization offers a lightweight and scalable solution. Inspired by it, we propose a novel GNSS-free collaborative perception framework that uses a lightweight regression-based LiDAR localization module to transmit pose and confidence representations, thereby reducing bandwidth, latency, and alignment failures (see Tab. 3).

However, several challenges remain. One major challenge is the lack of datasets that support both collaborative perception and regression-based LiDAR localization, which requires separate traversals of the same environment for training and testing. Another challenge lies in localization errors caused by imperfect pose estimation. Unlike previous methods (Hu et al. 2022; Lu et al. 2023; Zhang et al. 2024) that simulate localization noise using synthetic Gaussian perturbations to verify the robustness of their methods, our approach captures localization errors that naturally emerge from intricate and structured environmental conditions. As shown in Tab. 4 and the left figure of Fig. 2, detection performance differs between random noise and LiDAR localization errors, as random noise distributions fail to accurately model the structured and environment-driven characteristics of pose estimation errors.

Therefore, we propose a novel simulated dataset named V2VLoc. The V2VLoc dataset is composed of three subsets: Town1Loc, Town4Loc, and V2VDet. As illustrated in Fig. 4, we also develop a two-stage collaborative detection framework: (i) we train a single-agent Pose Generator with Confidence (PGC) model on the Town1Loc and Town4Loc subsets; and (ii) we perform per-frame pose and confidence estimation, followed by feature alignment and multi-agent collaborative detection on the V2VDet subset.

We also propose a novel module, Pose-Aware Spatio-Temporal Alignment Transformer (PASTAT), to alleviate the impact of localization errors. We first perform coarse alignment of features derived from various viewpoints, and incorporate the pose confidence as a learnable Confidence Embedding (CE). Subsequently, the coarsely matched features are further refined by the Feature Spatial Alignment (FSA) module for more accurate alignment. A Vision Transformer with Temporal Encoding (TE) is then applied to capture spatio-temporal context across multiple frames. Finally, the detection head processes the temporally encoded features to generate the final predictions.

To evaluate our method, we conduct extensive experiments on both V2VLoc and the real-world dataset V2V4Real (Xu et al. 2023). The results demonstrate that our method achieves the best performance in collaborative perception under GNSS-denied conditions, as shown in the right figure of Fig. 2, and also exhibits good performance in real-world scenarios. Our main contributions are summarized as follows:

- We are the first to apply LiDAR localization to solve the feature alignment problem in collaborative detection without GNSS signals. We establish a new perception pipeline that supports various fusion strategies and advances research on GNSS-free collaborative perception.
- We propose V2VLoc, a novel simulation dataset that supports both regression-based LiDAR localization and multi-agent collaborative object detection tasks, which lays a solid data foundation for collaborative perception in GNSS-denied scenarios.
- We design two key modules: the Pose Generator with Confidence (PGC) module and the Pose-Aware Spatio-Temporal Alignment Transformer (PASTAT) module to alleviate the impact of localization errors. Our method achieves superior performance across both the V2VLoc dataset and real-world benchmarks.

## Related Work

**Multi-Agent Collaborative Perception.** Multi-agent collaborative perception (Wang et al. 2020; Li et al. 2021; Xu et al. 2022c,a; Lu et al. 2023; Hu et al. 2023, 2024; Zhang et al. 2024; Song et al. 2025; Xia et al. 2025) has attracted increasing attention due to its potential to enhance the performance of individual agents significantly. By sharing sensory information among agents and transforming their observations into a unified coordinate system via ego poses, the field of view is broadened, effectively alleviating issues such as occlusion. However, most existing methods rely on external localization devices (e.g., GNSS or RTK) to obtain ego poses for cooperation, leading to degraded performance in GNSS-denied environments. FreeAlign (Lei et al. 2024) addresses this issue by estimating relative poses through graph matching. However, it heavily depends on the availability of multiple co-viewed objects.

**LiDAR Localization.** Map-based LiDAR localization (Choy, Park, and Koltun 2019; Xia et al. 2021; Zhang et al. 2023) estimates poses by matching query points with a prebuilt 3D map; however, it often incurs high storage and communication costs. To address the limitations of map-based localization, regression-based methods such as Absolute Pose Regression (APR) (Wang et al. 2021, 2023; Li et al. 2024a) and Scene Coordinate Regression (SCR) (Li et al. 2023; Yang et al. 2024a; Li et al. 2025; Chen et al. 2025) have emerged. APR directly regresses global poses from scene features, while SCR estimates poses by

| Dataset | Year | R/S | Sensor | V2X | Image (360°) | Agent Number | 3D Boxes | Multi-traversal |
|---|---|---|---|---|---|---|---|---|
| OPV2V | 2022 | S | L&C | V2V | ✓ | 7 | ✓ | × |
| V2X-Sim | 2022 | S | L&C | V2V&I | ✓ | 10 | ✓ | × |
| V2XSet | 2022 | S | L&C | V2V&I | ✓ | 5 | ✓ | × |
| DAIR-V2X | 2022 | R | L&C | V2I | × | 2 | ✓ | × |
| V2V4Real | 2023 | R | L&C | V2V | × | 2 | ✓ | × |
| V2X-Seq | 2023 | R | L&C | V2I | × | 2 | ✓ | × |
| V2XReal | 2024 | R | L&C | V2V&I | ✓ | 4 | ✓ | × |
| V2X-Radar | 2024 | R | L&C&R | V&I | × | 2 | ✓ | × |
| Open Mars | 2024 | R | L&C | V2V | × | 3 | × | ✓ |
| V2X-R | 2025 | S | L&C&R | V2V&I | ✓ | 5 | ✓ | × |
| V2VLoc (Ours) | 2025 | S | L&C&R | V2V | ✓ | 4 | ✓ | ✓ |
| - Town1Loc | 2025 | S | L&C&R | × | ✓ | 1 | ✓ | ✓ |
| - Town4Loc | 2025 | S | L&C&R | × | ✓ | 1 | ✓ | ✓ |
| - V2VDet | 2025 | S | L&C&R | V2V | ✓ | 4 | ✓ | × |

Table 1: Comparison of different datasets including OPV2V (Xu et al. 2022c), V2X-sim (Li et al. 2022), V2XSet (Xu et al. 2022b), DAIR-V2X (Yu et al. 2022), V2V4Real (Xu et al. 2023), V2X-Seq (Yu et al. 2023a), V2XReal (Xiang et al. 2024), V2X-Radar (Yang et al. 2024b), Open Mars (Li et al. 2024b), V2X-R (Huang et al. 2025) and V2VLoc. R/S: Real-world/Simulated, C: Camera, L: Lidar, R: Radar, V2V: vehicle-to-vehicle, V2I: Vehicle-to-infrastructure.

predicting LiDAR-to-world correspondences and applying RANSAC (Fischler and Bolles 1981). By eliminating the reliance on explicit maps, these methods provide a lightweight and flexible alternative, making them suitable for GNSS-denied scenarios in collaborative perception systems.

**Collaborative Perception Dataset.** Multi-agent collaborative perception has received increasing attention in recent years, with several studies focusing on the exploration and collection of more diverse datasets to support various tasks. For example, V2V4Real (Xu et al. 2023) is the first large-scale V2V dataset captured in real-world driving scenarios, while RCooper (Hao et al. 2024) introduces the pioneering large-scale dataset for roadside collaborative perception. V2X-R (Huang et al. 2025) is the first simulated V2X dataset that integrates LiDAR, camera, and 4D radar modalities. However, existing collaborative perception datasets are exclusively collected from single-pass trajectories, rendering them unsuitable for regression-based LiDAR localization tasks. This limitation presents a significant challenge for studying model robustness under GNSS-denied conditions.

## V2VLoc Dataset

### Overview of V2VLoc

Since the training and testing of regression-based LiDAR localization methods require different traversals of the same environment, this condition is not satisfied in the existing collaborative detection dataset. Therefore, we propose V2VLoc, the first dataset that supports both regression-based LiDAR localization and detection tasks. As shown in Tab. 1, V2VLoc is a comprehensive dataset consisting of three subsets: Town1Loc, Town4Loc, and V2VDet. All subsets include multimodal sensors, such as LiDAR, camera, and 4D radar. The dataset offers multi-traversal scans and 3D bounding box annotations for all frames, supporting both localization and detection tasks.

Figure 3: Trajectory map of Town1Loc and Town4Loc. The yellow star indicates the starting point of the traversal.

### Data Collection

The data collection for V2VLoc is performed using OpenCDA (Xu et al. 2021), a collaborative simulation platform based on CARLA (Dosovitskiy et al. 2017). This platform facilitates the simulation of collaborative driving scenarios involving multiple agents. Town1Loc and Town4Loc are constructed from the Town1 and Town4 maps of CARLA using a single agent for localization tasks. We visualize the agent's trajectory, as shown in Fig. 3, where Town1Loc and Town4Loc cover 31.31 km of roads that include underbridges, interchanges, roundabouts, ramps, gas stations, urban streets, and so on. V2VDet provides a variety of collaborative driving scenarios on Town1 and Town4 maps involving 2 to 4 agents per scene. The subset comprises 11,598 LiDAR and 4D-radar frames, 46,392 camera images, and 260,210 annotated vehicle 3D bounding boxes.

### Sensor Setup

Each vehicle in the V2VLoc dataset is equipped with multiple sensors to ensure comprehensive perception. The sensor

suite includes:

- Cameras: Four RGB cameras are mounted on each vehicle, positioned at (2.5, 0.1, 0.0), (0.0, 0.3, 1.8, 100), (0.0, –0.3, 1.8, –100), and (–2.0, 0.0, 1.5, 180), providing full 360° visual coverage around the vehicle.
- LiDAR: A single 64-channel LiDAR sensor is used, with a maximum range of 120 meters and a vertical field of view from –25° to 2°. It operates at a rotation frequency of 20 Hz and has a noise standard deviation of 0.02.
- 4D Radar: The radar has a sensing range of 150 meters, a 120° horizontal FOV, and a 30° vertical FOV, offering complementary motion and depth information under various weather and lighting conditions.
- GPS and IMU: The vehicle GNSS has an altitude noise of 0.001 meters. The vehicle IMU includes heading noise of 0.1°, and speed noise of 0.2 m/s. The RSU (Road Side Unit) GNSS provides altitude measurements with a noise level of 0.05 meters.

# Method

## Problem Formulation

In GNSS-denied environments, agents are unable to acquire accurate pose information via external localization systems. Therefore, unlike traditional collaborative detection frameworks, for each agent $i$, our collaborative detection process proceeds as follows:

$$\tau_i, \sigma_i = \mathcal{P}\left(O_i\right) \tag{1}$$

$$F_i^* = \Psi \cdot \mathcal{S}(\sigma) \cdot \mathcal{A}\left(f\left(O_i\right), f\left(\{O_j\}_{j \in \mathcal{N}}\right), \tau\right) \tag{2}$$

$$\hat{B}_i = D\left(F_i^*\right) \tag{3}$$

In Eq. (1), each agent first estimates its own pose $\tau_i$ and confidence $\sigma_i$ from the local observation $O_i$ using the pose estimation module $\mathcal{P}$. In Eq. (2), the feature extraction function $f(\cdot)$ encodes the local point cloud $O_i$ into features $f(O_i)$. These features are then coarsely aligned with the encoded features of neighboring agents $f\left(\{O_j\}_{j \in \mathcal{N}}\right)$ using the coarse alignment module $\mathcal{A}$ by $\tau$. Subsequently, a confidence-aware spatial alignment module $\mathcal{S}$ and a temporal transformer encoder $\Psi$ are applied to yield the aggregated feature $F_i^*$. Finally, In Eq. (3), the detection head $D$ takes $F_i^*$ as input to produce the final detection results $\hat{B}_i$.

## Pose Generator with Confidence (PGC)

To effectively overcome the limitation of the graph matching methods (Besl and McKay 1992; Lei et al. 2024) that require agents to share multiple commonly observed objects, a naive idea is to allow each agent to directly learn the correspondence between its observations and its pose by capturing the underlying geometric and semantic structure of the scene.

Inspired by this idea, we adopt the architecture of regression-based LiDAR localization, which estimates the 6-DoF pose by learning to transform the raw LiDAR point cloud (RPC) into the world-frame point cloud (WPC) and matching points between them using RANSAC (Fischler and Bolles 1981). However, as shown in the left figure of Fig. 2, regression-based LiDAR localization still suffers from pose estimation errors. Therefore, we propose the PGC module, which jointly estimates the pose and its confidence to enable more reliable collaborative alignment.

Specifically, we learn a mapping from the input raw point cloud to global scene coordinates $y_i$, the network regresses $y_i$ by minimizing the average Euclidean distance to the ground truth $y_i^*$, as shown in Eq. (4).

$$u_i = ||y_i - y_i^*||_1 \tag{4}$$

Meanwhile, we further propose to let the network learn to predict the pose error $\varepsilon$. The overall loss function is defined as follows:

$$\mathcal{L}_{reg} = \frac{1}{|\mathcal{Y}|} \sum_{y_i \in \mathcal{Y}} u_i + ||u_i - \varepsilon_i||_1 \tag{5}$$

where $\mathcal{Y}$ is the set of samples, and $|\mathcal{Y}|$ is the number of samples. And the pose confidence is derived from Eq. (6):

$$\sigma_i = \frac{1}{1 + \varepsilon_i^2} \tag{6}$$

A lower pose error results in a higher confidence, and a higher pose error results in a lower confidence, ensuring that more reliable pose estimates contribute more in downstream tasks such as feature alignment or fusion. Then, RANSAC (Fischler and Bolles 1981) is used to obtain the final poses and confidences.

Moreover, to accelerate network training, we adopt the Redundant Sample Downsampling (RSD) strategy (Li et al. 2025), which effectively reduces the number of redundant points on the input. This not only reduces the model parameter count but also enhances the training efficiency by focusing computation on informative samples.

## Pose-Aware Spatio-Temporal Alignment Transformer (PASTAT)

PASTAT consists mainly of the following components: (i) Feature Extraction and Coarse Alignment; (ii) Feature Spatial Alignment (FSA); (iii) Vision Transformer with Temporal Encoding; (iv) Decoder and Loss.

**Feature Extraction and Coarse Alignment.** Each agent $i$ receives a raw observation of point cloud $O_i$ as input. We use a shared feature encoder to extract semantically meaningful spatial features. With PGC, each agent $j \in \mathcal{N}$ transmits its encoded features $F_j$, the estimated pose $\tau_j \in \mathrm{SE}(3)$, and the corresponding confidence score $\sigma_j \in [0, 1]$ to the ego agent. With estimated poses $\tau$, we first compute the relative transformation from neighbor $j$ to ego agent $i$ and apply a rigid transformation to each neighbor's feature map to concatenate all features. Note that the features here are not fully aligned due to posture errors and have not been fused.

**Feature Spatial Alignment (FSA).** Although the coarsely aligned features provide a basic foundation for multi-agent fusion, they may still suffer from global pose noise, leading to sub-optimal detection performance. To address this limitation, we introduce the FSA module to enable precise feature-level alignment.

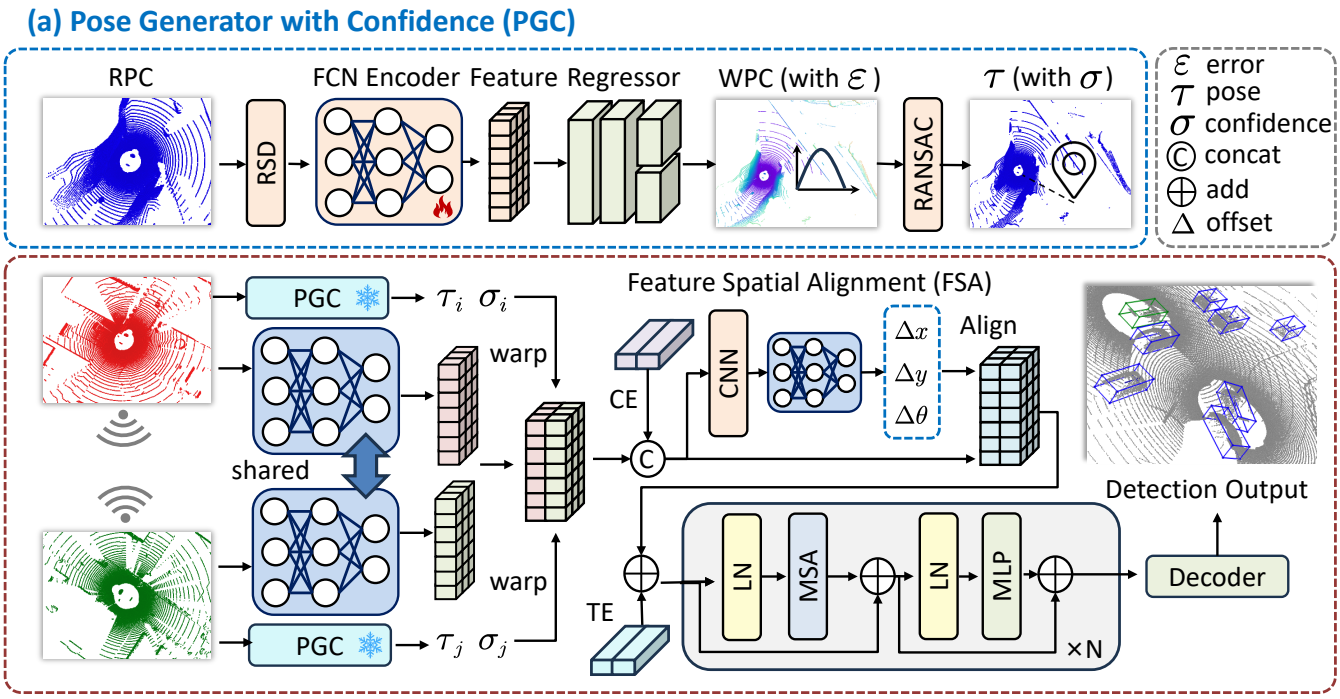


Figure 4: The overall architecture of the proposed framework. The architecture consists of two modules: (a) Pose Generator with Confidence (PGC), (b) Pose-Aware Spatio-Temporal Alignment Transformer (PASTAT). The agent obtains the global pose and confidence through PGC, and then aligns the features through PASTAT to obtain the collaborative detection result. *RSD*: Redundant Sample Downsampling; *RPC*: Raw LiDAR Point Cloud; *WPC*: Point Cloud in World coordinate system; *CE*: Confidence Embedding; *TE*: Temporal Encoding.

The core idea of FSA is to design a learnable alignment network that estimates inter-agent feature misalignments. Specifically, we first apply a normalization operation to the confidence scores $\sigma$ generated by the PGC module, referred to as Confidence Embedding (CE), and then concatenate the normalized confidence with the extracted feature maps, as shown in Eq. (7). This enables the network to explicitly incorporate pose reliability into the alignment process.

$$F_i' = F_i \oplus \frac{\sigma_i}{\sum_j^N \sigma_j} \tag{7}$$

To achieve accurate alignment, we introduce a dedicated alignment network to predict spatial transformation parameters $\Delta T_{j\to i} \in \mathbb{R}^K$, which is defined with $K$ degrees of freedom. In our design, we adopt a 3-DoF transformation, where the predicted offset includes translation in the $x$ and $y$ directions, and a rotation angle $\theta$. Our network uses convolutional neural networks (CNNs) and fully connected layers to learn the offsets of the feature shift between agents. These offsets are then used to adjust the spatial positions of the fused features accordingly.

**Vision Transformer with Temporal Encoding.** We adopt a Vision Transformer (ViT) encoder (Xu et al. 2022b) for global context modeling, treating the aligned feature map as token sequences. Each transformer layer comprises Multi-Head Self-Attention (MSA) and a Feed-Forward Network (MLP), both with Layer Normalization (LN) and residual connections. Given input $z^{(l-1)}$ to the $l^{th}$ layer, the computations are defined as:

$$z^{(l)'} = \text{MSA}\left(\text{LN}\left(z^{(l-1)}\right)\right) + z^{(l-1)} \tag{8}$$

$$z^{(l)} = \text{MLP}\left(\text{LN}\left(z^{(l)'}\right)\right) + z^{(l)'} \tag{9}$$

To enable temporal awareness, we also incorporate a temporal encoding scheme (Xu et al. 2022b) into the input of the transformer. Concretely, given a sequence of aligned features from multiple time steps $\{\tilde{F}_i^{(t)}\}_{t=1}^T$, we inject a learnable or sinusoidal temporal encoding $E_t$ into each feature token before feeding it into the Transformer:

$$z_t^{(0)} = \text{Flatten}(\tilde{F}_i^{(t)}) + E_t \tag{10}$$

where $E_t \in \mathbb{R}^D$ represents the temporal encoding corresponding to the time step $t \in \{1, 2, \ldots, T\}$, and $D$ is the dimension of the embedding of the token. For each time step $t$ , the temporal encoding vector $E_t$ is defined as:

$$E_t^{(2k)} = \sin\left(\frac{t}{10000^{(2k)/D}}\right) \tag{11}$$

$$E_t^{(2k+1)} = \cos\left(\frac{t}{10000^{(2k+1)/D}}\right) \tag{12}$$

where $k = 0, 1, \ldots, D/2 - 1$.

| Method | Reference | V2V4Real | | | V2VDet | | |
|---|---|---|---|---|---|---|---|
| | | AP@0.3 | AP@0.5 | AP@0.7 | AP@0.3 | AP@0.5 | AP@0.7 |
| No Fusion | - | 47.50 | 39.83 | 22.02 | 40.16 | 37.28 | 22.31 |
| Late Fusion | - | 40.18 | 34.60 | 15.64 | 50.88 | 47.53 | 29.58 |
| Where2comm (Hu et al. 2022) | NeurIPS2022 | 61.30 | 57.61 | 37.75 | 57.80 | 49.38 | 33.95 |
| CoBEVT (Xu et al. 2022a) | CORL2022 | 59.03 | 56.11 | 34.69 | 63.68 | 59.50 | 39.11 |
| V2X-ViT (Xu et al. 2022b) | ECCV2022 | 60.15 | 56.90 | 35.84 | 67.52 | 62.70 | 43.46 |
| CoAlign (Lu et al. 2023) | ICRA2023 | 62.11 | 58.93 | 34.38 | 66.75 | 62.98 | 37.50 |
| ERMVP (Zhang et al. 2024) | CVPR2024 | 60.86 | 58.90 | 38.74 | 67.86 | 64.24 | 47.13 |
| TraF-Align (Song et al. 2025) | CVPR2025 | 62.11 | 56.11 | 31.54 | 72.65 | 66.75 | 48.29 |
| PASTAT (Ours) | - | **63.52** | **61.51** | **40.29** | **76.97** | **71.15** | **52.55** |

Table 2: Performance comparison of vehicle class on V2V4Real (Xu et al. 2023) test and V2VDet test dataset. The results are reported in AP@0.3 / 0.5 / 0.7, and the agent uses the ground truth pose with 1.0 / 1.0 noise level (m/°) in V2V4Real (Xu et al. 2023), while PGC is used to obtain the pose in V2VDet.

| Method | Time (s) | $\mathcal{C}$ (log2) | $\delta_s$ (%) |
|---|---|---|---|
| ICP (Besl and McKay 1992) | 0.6009 | 7.37 | 0.813 |
| FreeAlign (Lei et al. 2024) | 0.0878 | | 60.26 |
| HypLiLoc (Wang et al. 2023) | 0.5652 | 4.62 | 80.96 |
| SGLoc (Li et al. 2023) | 0.0242 | | 83.47 |
| DiffLoc (Li et al. 2024a) | 0.1911 | | 81.32 |
| LightLoc (Li et al. 2025) | 0.0081 | | 98.35 |

Table 3: Comparison of Alignment Performance. We compared alignment results in the V2VDet dataset using ground truth bounding boxes for ICP and FreeAlign. Time represents the average alignment time, $\mathcal{C}$ represents the communication volume during alignment, $\delta_s$ (%) indicates the success rate, which means translation error is less than 3m.

**Decoder and Loss.** Based on the final fused feature representation, we generate the detection outputs using a detection decoder. Each predicted bounding box $\hat{b}_i$ represents a rotated 3D object, parameterized as: $\hat{b}_i = (x, y, z, h, w, l, \theta)$. Following prior work (Lang et al. 2019), we adopt the Smooth $L1$ loss for bounding box regression and the Focal Loss (Lin et al. 2020) for classification, which together ensure robust training under class imbalance and spatial uncertainty.

# Experiments

## Dataset and Evaluation Metrics

We conduct experiments on both our V2VLoc dataset and the real-world V2V4Real (Xu et al. 2023) dataset. For V2VLoc, we use 6,697 training, 2,017 validation, and 2,884 test frames for collaborative 3D object detection. For V2V4Real (Xu et al. 2023), it is the first large-scale real-world dataset for vehicle-to-vehicle (V2V) collaborative perception. It covers 410 kilometers of driving, providing over 20K frames of LiDAR data, with 14,210 frames for training, 2,000 for validation, and 3,986 for testing. To ensure fair and consistent evaluation, we adopt the official metrics used in previous works, including mean Average Precision (mAP) under IoU thresholds of 0.3, 0.5, and 0.7.

## Implementation Details

For the PGC module, we implement our method using PyTorch and train the model on a multi-GPU server with 24 data loading workers. The model is optimized using the AdamW optimizer with an initial learning rate of 1e-3 and a weight decay of 1200. We adopt a total of 100 training epochs with a batch size of 100.

For the PASTAT module, we train the network on the V2VDet dataset using the OpenCOOD (Xu et al. 2022c) framework and adopt PointPillar (Lang et al. 2019) as the lightweight backbone model. Models are trained for 60 epochs with a batch size of 2 using the Adam optimizer (initial LR = 0.001, weight decay=1e-4) and a MultiStep scheduler (decay at epochs 15 and 50). All experiments were trained on 4 NVIDIA GeForce RTX 3090 GPUs.

## Quantitative Evaluation

**Comparison of Alignment Performance.** As shown in Tab. 3, regression-based localization methods clearly outperform traditional approaches. LightLoc (Li et al. 2025) achieves the highest alignment success rate, 98.35%, with a low communication volume of 4.62 and a fast runtime of 0.0081s, which demonstrates the effectiveness of LiDAR localization methods for collaborative feature alignment in GNSS-denied environments.

**Comparison of Detection Performance.** Tab. 2 presents the comparison of the 3D object detection performance of vehicle class between models in the V2V4Real (Xu et al. 2023) and V2VDet datasets.

For the V2VDet dataset, we simulate a GNSS-denied environment, where all fusion methods utilize the PGC module to generate poses. Our proposed PASTAT achieves the best performance, outperforming the previous state-of-the-art by 4.40% and 4.26% on AP@0.5 and AP@0.7, respectively.

For the V2V4Real (Xu et al. 2023) dataset, we do not train our PGC module on it, as it contains only a single LiDAR traversal per location. This setup violates a key

| Method/Metric | AP@0.3/0.5/0.7 | | | | |
|---|---|---|---|---|---|
| Noise Level $\sigma_t/\sigma_r$ (m/°) | 0.0/0.0 | 1.0/1.0 | 2.0/2.0 | 3.0/3.0 | 4.0/4.0 |
| No Fusion | 40.16/37.28/22.31 | | | | |
| Early Fusion | 62.95/61.05/48.43 | 50.02/44.59/27.05 | 48.81/46.55/32.65 | 50.60/48.50/34.60 | 51.33/49.21/35.25 |
| Late Fusion | 58.03/57.03/47.15 | 49.29/39.33/27.27 | 50.58/49.57/40.78 | 50.68/49.54/40.77 | 50.60/49.53/40.68 |
| CoBEVT | 65.12/62.59/43.86 | 61.21/57.22/37.71 | 58.30/55.54/37.94 | 57.95/56.03/39.17 | 59.24/57.48/40.91 |
| Where2comm | 68.77/59.89/36.64 | 66.11/57.78/35.35 | 63.29/55.83/34.45 | 63.77/57.59/36.49 | 65.36/59.35/38.05 |
| Coalign | 68.52/65.41/43.24 | 58.14/52.48/31.07 | 57.07/53.66/31.96 | 57.27/53.47/31.00 | 57.18/53.77/32.05 |
| V2X-ViT | 73.14/69.11/48.69 | 63.59/56.61/27.56 | 61.13/56.74/27.78 | 61.30/57.76/29.71 | 61.29/57.93/30.14 |
| ERMVP | 69.95/69.00/57.93 | 55.50/50.64/27.28 | 45.77/40.73/30.32 | 42.41/41.60/32.88 | 42.69/42.10/33.84 |
| Traf-Align | 78.24/73.08/42.38 | 72.19/59.51/36.26 | 65.82/56.08/31.60 | 62.98/55.23/32.30 | 62.49/56.86/33.90 |
| PASTAT(Ours) | **76.97/71.15/52.55** | | | | |

Table 4: Performance comparison of different models on the V2VDet dataset under different GNSS perturbations. We evaluate AP@0.3/0.5/0.7 of different fusion methods, and our modules consistently maintain strong performance as they are not affected by GNSS disturbances.

assumption of regression-based LiDAR localization: training and testing must occur on distinct traversals to prevent overfitting to scene-specific geometry. Therefore, to fairly benchmark downstream collaborative detection performance, ground truth poses with 1.0 / 1.0 noise level (m/°) (Hu et al. 2022; Zhang et al. 2024; Lei et al. 2024) are used, with pose confidence obtained from pose error. It can be seen that the PASTAT module can effectively solve the problem of inaccurate LiDAR pose in both real and simulated datasets through feature alignment.

**Robustness to Pose Errors by GNSS.** To assess robustness under GNSS-denied conditions, we simulate varying GNSS noise levels and evaluate their impact on different collaborative detection pipelines. Specifically, we simulate varying levels of localization noise by adjusting the standard deviation of Gaussian pose perturbations, ranging from 0.0/0.0 to 4.0/4.0 (m/°). The performance of different models under these conditions is evaluated in terms of AP@0.3/0.5/0.7 (see Tab. 4). While all models exhibit a noticeable decline in performance as the noise level increases, the degree of robustness varies significantly among them. Notably, our PASTAT model, which incorporates pose generation via the PGC module, demonstrates superior robustness to GNSS localization perturbations.

## Ablation Studies

To evaluate the individual contributions of each component in our proposed framework, we conduct a comprehensive ablation study on the V2VLoc dataset. The ablation focuses on four key modules: Pose Generator with Confidence (PGC), Confidence Embedding (CE), Feature Spatial Alignment (FSA), and Temporal Encoding (TE). The corresponding experimental results are summarized in Tab. 5. We take the PointPillars (Lang et al. 2019) and Vision Transformer Encoding (Xu et al. 2022b) as our baseline and progressively add the following modules to evaluate the contribution of each component: (i) PGC, (ii) CE, (iii) FSA, and (iv) TE. The FSA does not include the CE module in this context. As shown in Tab. 5, all modules contribute positively to performance improvement. In particular, the proposed PGC module improves AP@0.3/0.5/0.7 by 15.42%/14.19%/15.42%, respectively, while the PASTAT module further boosts AP@0.3/0.5/0.7 by 19.02%/18.47%/12.86%, respectively.

| ID | PGC | PASTAT | | | V2VLoc | | |
|---|---|---|---|---|---|---|---|
| | | CE | FSA | TE | AP@0.3 | AP@0.5 | AP@0.7 |
| 1 | | | | | 42.53 | 38.49 | 24.27 |
| 2 | ✓ | | | | 57.95 | 52.68 | 39.69 |
| 3 | ✓ | ✓ | | | 69.43 | 63.43 | 41.07 |
| 4 | ✓ | ✓ | ✓ | | 72.82 | 66.70 | 48.19 |
| 5 | ✓ | ✓ | ✓ | ✓ | **76.97** | **71.15** | **52.55** |

Table 5: Results of the ablation study of the core components on the V2VLoc dataset. We use PointPillars (Lang et al. 2019) and ViT (Xu et al. 2022b) as baselines and gradually add our modules to verify the performance of our method.

## Conclusion

In this paper, we propose a novel GNSS-free collaborative perception framework that enables robust multi-agent detection without GNSS signals. By leveraging LiDAR localization, our method establishes accurate 3D correspondences between local observations and a shared global scene, allowing each agent to estimate its pose without relying on external localization systems. We propose a new dataset, V2VLoc, and to enhance the robustness and accuracy of the fusion process, we further incorporate modules such as PGC and PASTAT. Extensive experiments on the V2VLoc and V2V4Real demonstrate the effectiveness of our approach.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.42571514).

## References


Besl, P.; and McKay, N. D. 1992. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2): 239–256.

Chen, Y.; Li, Q.; Yang, Y.; Li, W.; Ao, S.; and Wang, C. 2025. Unleashing the Power of Data Generation in One-Pass Outdoor LiDAR Localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 199–208.

Choy, C.; Park, J.; and Koltun, V. 2019. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8958–8966.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Hao, R.; Fan, S.; Dai, Y.; Zhang, Z.; Li, C.; Wang, Y.; Yu, H.; Yang, W.; Jirui, Y.; and Nie, Z. 2024. RCooper: A Real-world Large-scale Dataset for Roadside Cooperative Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22347–22357.

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

Hu, Y.; Lu, Y.; Xu, R.; Xie, W.; Chen, S.; and Wang, Y. 2023. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9243–9252.

Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024. Communication-Efficient Collaborative Perception via Information Filling with Codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15481–15490.

Huang, X.; Wang, J.; Xia, Q.; Chen, S.; Yang, B.; Li, X.; Wang, C.; and Wen, C. 2025. V2X-R: Cooperative LiDAR-4D Radar Fusion with Denoising Diffusion for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27390–27400.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 12697–12705.

Lei, Z.; Ni, Z.; Han, R.; Tang, S.; Feng, C.; Chen, S.; and Wang, Y. 2024. Robust collaborative perception without external localization and clock devices. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 7280–7286. IEEE.

Li, W.; Liu, C.; Yu, S.; Liu, D.; Zhou, Y.; Shen, S.; Wen, C.; and Wang, C. 2025. LightLoc: Learning Outdoor LiDAR Localization at Light Speed. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6680–6689.

Li, W.; Yang, Y.; Yu, S.; Hu, G.; Wen, C.; Cheng, M.; and Wang, C. 2024a. Diffloc: Diffusion model for outdoor lidar localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15045–15054.

Li, W.; Yu, S.; Wang, C.; Hu, G.; Shen, S.; and Wen, C. 2023. SGLoc: Scene geometry encoding for outdoor LiDAR localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9286–9295.

Li, Y.; An, Z.; Wang, Z.; Zhong, Y.; Chen, S.; and Feng, C. 2022. V2x-sim: A virtual collaborative perception dataset for autonomous driving. *arXiv preprint arXiv:2202.08449*.

Li, Y.; Li, Z.; Chen, N.; Gong, M.; Lyu, Z.; Wang, Z.; Jiang, P.; and Feng, C. 2024b. Multiagent multitraversal multimodal self-driving: Open mars dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22041–22051.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34: 29541–29552.

Lin, T.-Y.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42: 318–327.

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

Song, Z.; Yang, L.; Wen, F.; and Li, J. 2025. Traf-align: Trajectory-aware feature alignment for asynchronous multi-agent perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12048–12057.

Wang, S.; Kang, Q.; She, R.; Wang, W.; Zhao, K.; Song, Y.; and Tay, W. P. 2023. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5176–5185.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. 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*, 605–621. Springer.

Wang, W.; Wang, B.; Zhao, P.; Chen, C.; Clark, R.; Yang, B.; Markham, A.; and Trigoni, N. 2021. Pointloc: Deep pose regressor for lidar point cloud localization. *IEEE Sensors Journal*, 22(1): 959–968.

Xia, Q.; Lin, W.; Xiang, H.; Huang, X.; Chen, S.; Dong, Z.; Wang, C.; and Wen, C. 2025. Learning to Detect Objects from Multi-Agent LiDAR Scans without Manual Labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1418–1428.

Xia, Y.; Xu, Y.; Li, S.; Wang, R.; Du, J.; Cremers, D.; and Stilla, U. 2021. SOE-Net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11348–11357.

Xiang, H.; Zheng, Z.; Xia, X.; Xu, R.; Gao, L.; Zhou, Z.; Han, X.; Ji, X.; Li, M.; Meng, Z.; et al. 2024. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision*, 455–470. Springer.

Xu, R.; Guo, Y.; Han, X.; Xia, X.; Xiang, H.; and Ma, J. 2021. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1155–1162. IEEE.

Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; and Ma, J. 2022a. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*.

Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; et al. 2023. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13712–13722.

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

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022c. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, 2583–2589. IEEE.

Yang, B.; Li, Z.; Li, W.; Cai, Z.; Wen, C.; Zang, Y.; Muller, M.; and Wang, C. 2024a. Lisa: Lidar localization with semantic awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15271–15280.

Yang, L.; Zhang, X.; Wang, C.; Li, J.; Ma, J.; Song, Z.; Zhao, T.; Song, Z.; Wang, L.; Zhou, M.; et al. 2024b. V2X-radar: A multi-modal dataset with 4D radar for cooperative perception. *arXiv preprint arXiv:2411.10962*.

Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; et al. 2022. 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*, 21361–21370.

Yu, H.; Yang, W.; Ruan, H.; Yang, Z.; Tang, Y.; Gao, X.; Hao, X.; Shi, Y.; Pan, Y.; Sun, N.; Song, J.; Yuan, J.; Luo, P.; and Nie, Z. 2023a. 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*.

Yu, S.; Sun, X.; Li, W.; Wen, C.; Yang, Y.; Si, B.; Hu, G.; and Wang, C. 2023b. Nidaloc: Neurobiologically inspired deep lidar localization. *IEEE Transactions on Intelligent Transportation Systems*, 25(5): 4278–4289.

Zhang, J.; Yang, K.; Wang, Y.; Wang, H.; Sun, P.; and Song, L. 2024. ERMVP: Communication-Efficient and Collaboration-Robust Multi-Vehicle Perception in Challenging Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12575–12584.

Zhang, X.; Yang, J.; Zhang, S.; and Zhang, Y. 2023. 3D registration with maximal cliques. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17745–17754.