Noise is a Good Teacher: A Noise-Driven Framework for Robust Collaborative Perception

Chengyan Huang[1], Zhongxiang Zhao[1], Ziyao Zhang[1], Zihao Yang[1], and Lin Wang[1]*

School of Artificial Intelligence (School of Software), Yanshan University, Qinhuangdao 066004, China
wangllinn@gmail.com

Abstract. Collaborative perception mitigates occlusions and extends the perception range, yet its accuracy remains susceptible to both pose and feature noise. Given these limitations, we propose **NoiseGT**, a noise-driven framework built upon the principle that "Noise is a Good Teacher". Instead of treating noise as a disturbance, NoiseGT leverages it as a self-supervised signal to enhance robustness at both the pose and feature levels. Specifically, a pose calibration module first applies multi-level noise injection, followed by graph-based alignment refinement. This process enables the module to learn pose-invariant alignment despite localization uncertainties. Meanwhile, a feature denoising module adopts a "high-noise-over-low-noise" strategy to inject high-magnitude noise into bird's-eye-view (BEV) features, guiding the denoiser toward noise-robust representations without relying on clean intermediate supervision. Together, these components form a noise-driven learning process that promotes consistent fusion across agents. At the most challenging noise setting ($(\sigma_t, \sigma_r) = (1.2\,\mathrm{m}, 1.2^\circ)$), NoiseGT achieves state-of-the-art performance over previous methods, improving AP@0.5/0.7 by a relative **12.03%/14.56%** on V2X-Sim, **2.67%/0.89%** on DAIR-V2X, and **6.12%/9.69%** on OPV2V. These results highlight the effectiveness and generality of noise-driven learning for robust collaborative perception.

Keywords: BEV Denoising · Collaborative Perception · Self-Supervised Learning

1 Introduction

With the rapid development of autonomous driving, individual agents (*e.g.*, vehicles) have achieved strong perception capabilities [4, 5, 11, 12, 19, 20, 34, 37]. However, their performance remains limited in complex traffic scenarios due to occlusions and long-range perception challenges [22, 34]. To overcome these limitations, collaborative perception allows multiple agents to share and fuse their sensing information [11, 34]. Specifically, the fusion of Global Navigation Satellite System (GNSS) and Inertial Measurement Unit (IMU) data provides pose

* Corresponding author.

information, while multi-view cameras or LiDAR sensors capture environmental details. By sharing and fusing information among agents, collaborative perception significantly expands the perception range and enhances robustness.

High-quality collaborative perception relies on two essential factors: accurate pose information and stable BEV features. During collaboration, each agent must obtain the relative transformations based on precise poses and align the BEV features of other agents into its own coordinate system for feature fusion. However, in real-world scenarios, inherent GNSS localization errors and IMU drift introduce pose noise [2, 8, 10, 21]. Meanwhile, sensor imperfections and interpolation processes [43] during BEV feature extraction further introduce feature noise [30, 43]. These two types of noise exist not only independently but also interact to amplify one another during cross-agent fusion [10, 33]. Such interplay between misalignment and accumulated noise significantly degrades the reliability of collaborative perception.

In collaborative perception, existing studies have mainly focused on mitigating pose noise, while feature noise and its coupling with pose noise have been largely overlooked. **For pose noise in collaborative perception,** early-fusion methods [4] typically employ iterative closest point (ICP) [25] or other registration-based techniques [1, 41] for alignment. In contrast, intermediate- and late-fusion frameworks leverage graph-based optimization [13, 21] or matching-based refinement [10, 17, 28] to achieve more reliable cross-agent alignment. **For feature noise in single-agent BEV perception,** recent advances [12, 23, 39, 43] have introduced diffusion-based denoising frameworks that typically rely on explicit dense BEV intermediate targets (*e.g.*, semantic layout) and iterative sampling to reconstruct stabilized BEV features. However, these methods have not been extended to collaborative perception.

In practice, collaborative perception systems inevitably suffer from both pose and feature noise [8, 10, 13, 21, 28], which jointly degrade fusion quality and overall robustness. This motivates us to explore **how collaborative perception can remain robust when both forms of noise coexist (Q1)**.

While attempting to answer **Q1**, we find that most BEV feature denoising methods [31, 32, 39, 43] are designed to reconstruct or refine BEV features under the strong assumption that explicit dense intermediate supervision is available (*e.g.*, ground-truth BEV semantic layouts or depth distributions). In real-world collaborative perception, however, such clean intermediate targets are often difficult or even impossible to obtain due to sensor imperfections, environmental variations, and annotation cost [7, 15, 24]. This motivates a more fundamental question that must be addressed for practical robustness: **whether BEV feature denoising necessarily depends on clean intermediate supervision (Q2)**. Answering this question is crucial for scalable collaborative perception without extra dense intermediate annotations, especially under noisy pose and feature conditions.

This question motivates us to rethink the essence of denoising. Traditionally, noise is regarded as an undesirable disturbance, and the goal of a denoising model is to eliminate it as much as possible [9, 31, 32]. In contrast, we argue that the

essence of denoising lies not in reconstructing a noise-free target, but in learning meaningful representations of noise itself. This perspective leads us to further consider that if a model can learn to denoise from strong to normal noise levels (which implies that it has captured sufficient noise representations), it is likely capable of generalizing this process from normal noise to noise-free conditions as well. Based on this assumption, we present our core inference: **noise is not merely something to be removed, but can also serve as a good teacher**.

Building upon this inference, we propose **NoiseGT**, a noise-driven framework for robust collaborative perception that jointly mitigates pose and feature noise through two complementary modules. Specifically, the Pose Calibration Module employs multi-level noise injection and graph-based alignment refinement to achieve consistent calibration under localization uncertainty. Meanwhile, the Feature Denoising Module employs a lightweight design that transforms noise from a nuisance into a self-supervised learning signal. It follows a "high-noise-over-low-noise" strategy, where stronger artificial noise suppresses weaker perturbations and encourages the model to learn consistency under disturbances. By continuously confronting injected noise during training, the network learns to maintain stable and task-relevant features without requiring explicit intermediate feature supervision.

Furthermore, the joint optimization of the self-supervised denoising objective and the downstream detection loss promotes the learning of features that are both stable and semantically aligned with the perception task. In essence, **Noise is a Good Teacher** guides the model toward stability, consistency, and robustness, thereby enabling reliable multi-agent fusion under noisy conditions.

In summary, our main contributions are as follows:

- We present, to the best of our knowledge, the first systematic attempt to explicitly formulate and jointly address pose noise, BEV feature noise, and their coupling in collaborative perception to answer **Q1**.
- We propose the **Noise is a Good Teacher** principle for collaborative perception to answer **Q2**. This principle treats noise as a self-supervised signal rather than a disturbance, guiding the model toward stability and consistency under uncertainty.
- We establish a noise-driven framework that systematically addresses both pose and feature uncertainty. The Pose Calibration Module employs multi-level noise injection to improve generalization under localization errors, while the Feature Denoising Module leverages strong artificial noise to stabilize self-supervised denoising and enhance robustness.
- We demonstrate that our method achieves state-of-the-art performance on both simulated collaborative perception datasets (OPV2V and V2X-Sim) [18, 37] and the real-world dataset (DAIR-V2X) [40]. While some baselines may perform slightly better under noise-free conditions, NoiseGT exhibits remarkable robustness under high pose noise settings and maintains strong overall performance across all benchmarks.

2 Related Work

Collaborative Perception. Collaborative perception mitigates inherent limitations of single-agent sensing, such as occlusions and long-range challenges [19, 20, 34, 37], while enhancing the ego agent (Ego) features with information from other connected and autonomous vehicles (CAVs). The emergence of several recent datasets has greatly advanced research in collaborative perception. Specifically, OPV2V [37] and V2X-Sim [18] provide large-scale simulation environments, while DAIR-V2X [40] offers real-world perception data. Existing approaches can be categorized as early, intermediate, or late fusion: early fusion [4] transmits raw sensor data with high bandwidth cost, late fusion [29] shares only detection results but is noise-sensitive, and intermediate fusion [11, 18–21, 37, 40] exchanges features, balancing bandwidth and accuracy. However, most fusion strategies assume clean and perfectly aligned features, an assumption that is often violated in practice due to pose errors and sensor noise. These misalignments have motivated a line of research on pose correction and feature refinement.

Pose Correction. Regardless of the fusion strategy, accurate pose information is essential for aligning and fusing the cooperative data received from CAVs. In practice, however, pose estimation is inevitably affected by noise. Several works [13, 21] have explored pose correction using graph-based or mathematical approaches. Classical iterative closest point (ICP) [25] aligns point clouds by iteratively matching points and optimizing the transformation. VIPS [26] models the relationships between bounding boxes as a graph and identifies correspondences via vertex and edge matching. CBM [29] reduces incorrect matches by enforcing consensus between global and local correspondences. GraphPS [17] employs a graph neural network to generate optimal matching sequences, improving matching success rates. Additionally, CoAlign [21] leverages an agent–object pose graph for pose correction, while RoCo [13] addresses pose inconsistencies via object matching and robust graph optimization.

Feature Refinement. In collaborative perception, BEV features extracted from LiDAR or multi-sensor data are often corrupted by sensor noise, occlusions, and pose misalignment, which can degrade downstream perception performance. To address this, recent works often employ diffusion models, a class of generative models that gradually transform simple distributions into complex ones through iterative denoising [9, 12, 27]. Building on this, several works [12, 39, 43] have explored diffusion-based BEV feature denoising. For instance, DiffBEV [43] applies a conditional diffusion model using semantic features derived from depth distributions as guidance, while BEVDiffuser [39] leverages ground-truth object layouts, using object location and category to guide denoising. Although these methods are effective, they rely on idealized conditions and incur high computational costs, which motivates our research into lightweight, self-supervised BEV feature denoising.

3 Method

We present **NoiseGT**, a framework that transforms uncertainty into explicit supervision for robust collaborative perception. Guided by the principle that Noise is a Good Teacher, NoiseGT jointly addresses pose misalignment and feature corruption via two **plug-and-play** components: the Pose Calibration Module for distributionally robust calibration and the Feature Denoising Module for self-supervised BEV denoising. Together, they form a unified refinement pipeline that improves reliability under diverse noisy conditions.

3.1 Overall Framework

The standard intermediate-fusion pipeline for N collaborating agents can be formulated as follows. Each agent i processes its raw LiDAR observations $\mathbf{O}_i$ to extract BEV features:

$$\mathbf{F}_i = f_{\text{encoder}}(\mathbf{O}_i), \tag{1}$$

$$\mathbf{M}_{j\to i} = f_{\text{transform}}(\boldsymbol{\xi}_i, \mathbf{F}_j, \boldsymbol{\xi}_j), \tag{2}$$

$$\mathbf{F}'_i = f_{\text{fusion}}(\mathbf{F}_i, \{\mathbf{M}_{j\to i}\}_{j=1,2,\ldots,N}), \tag{3}$$

$$\mathbf{B}_i = f_{\text{decoder}}(\mathbf{F}'_i), \tag{4}$$

where $\boldsymbol{\xi}_i = (x_i, y_i, z_i, \theta_i, \phi_i, \psi_i)$ denotes the 6-d.o.f. pose of agent i, and features $\mathbf{F}_j$ from neighboring agents are transformed into the coordinate frame of agent i as messages $\mathbf{M}_{j\to i}$ based on their relative poses. The aggregated feature $\mathbf{F}'_i$ is then decoded by $f_{\text{decoder}}(\cdot)$ to produce the final output $\mathbf{B}_i$.

In real-world deployments, the assumptions of perfect poses and noise-free features rarely hold, making intermediate fusion particularly vulnerable to both misalignment and representation corruption. As shown in Fig. 1, NoiseGT addresses these two uncertainty sources in a unified manner via the Pose Calibration Module and the Feature Denoising Module.

3.2 Pose Calibration Module (PCM)

To enhance robustness against localization uncertainty, PCM employs a two-stage strategy that integrates multi-level noise injection for noise-aware training and graph-based alignment refinement for fine-grained correction.

Multi-level noise injection. Existing calibration pipelines are often trained with a **single** perturbation scale, which biases the model to a narrow error range and degrades robustness when localization noise varies across scenes, agents, or hardware. We instead aim for distributionally robust calibration under diverse uncertainties.

During training, we uniformly sample translational and rotational noise levels (σ_t, σ_r) from $\mathcal{S} = \{0.0, 0.2, 0.4, 0.6\}$, exposing the pipeline to a mixture of Gaussian perturbations at multiple scales and improving generalization to unseen noise magnitudes.

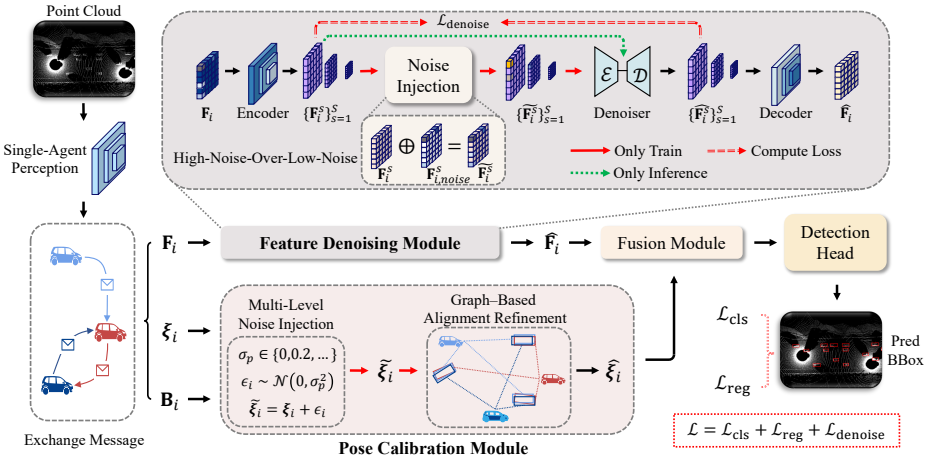


Fig. 1: Overview of the proposed NoiseGT framework. NoiseGT is a **noise-driven** and **plug-and-play** refinement pipeline for intermediate fusion. It systematically mitigates two major uncertainty sources in collaborative perception: pose noise and feature noise. The Pose Calibration Module calibrates noisy poses via distributionally robust training and consistency-based refinement, while the Feature Denoising Module learns self-supervised, cross-noise invariant BEV representations. These two modules act synergistically to improve robustness under diverse noisy conditions.

Although PCM is non-parametric, pose perturbations change the inputs to the full perception pipeline, thus the downstream learnable components are implicitly regularized to be insensitive to misalignment across perturbation levels.

This can be understood via local sensitivity analysis. Let $\epsilon \sim \mathcal{N}(0, \Sigma)$ be pose noise with block-diagonal covariance induced by (σ_t, σ_r). Then

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\big[\ell\big(g(\boldsymbol{\xi}+\epsilon)\big)\big] \approx \ell\big(g(\boldsymbol{\xi})\big) + \tfrac{1}{2}\,\mathrm{tr}\big(J^\top H_\ell J\,\Sigma\big), \tag{5}$$

where $J = \partial g/\partial \boldsymbol{\xi}$ and H_ℓ is the Hessian of ℓ. With multi-level noise injection, taking expectation over $\{\Sigma_k\}$ for $(\sigma_t, \sigma_r) \in \mathcal{S}$ yields

$$\mathbb{E}_k\,\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma_k)}\big[\ell\big(g(\boldsymbol{\xi}+\epsilon)\big)\big] \approx \ell\big(g(\boldsymbol{\xi})\big) + \tfrac{1}{2}\,\mathbb{E}_k\big[\,\mathrm{tr}(J^\top H_\ell J\,\Sigma_k)\big]. \tag{6}$$

This shows that multi-level noise injection optimizes an expectation over regularization strengths, promoting smooth calibration and distributional robustness.

Graph-based alignment refinement. Given perturbed poses $\tilde{\boldsymbol{\xi}}_i$, we refine them via an agent–object pose graph to enforce cross-agent consistency over co-observed objects, producing calibrated poses $\hat{\boldsymbol{\xi}}_i$ for Eq. (2). Our PCM is **solver-agnostic** and can be instantiated with any off-the-shelf pose-graph optimizer. We use the open-source solver in CoAlign [21] for reproducibility.

3.3 Feature Denoising Module (FDM)

Even with accurately calibrated poses, BEV features may still degrade due to sensor noise, discretization, or projection artifacts. FDM addresses this by casting unknown corruptions as a self-supervised denoising problem.

Rationale for fixed-level noise. The noise scheduling differs between the two modules. PCM employs multilevel perturbations to span diverse pose errors, whereas FDM adopts a fixed high noise level ($\sigma_f = 0.2$) under the high-noise-over-low-noise strategy. Injecting isotropic Gaussian noise defines a consistent perturbation envelope with sufficient magnitude, rather than assuming corruptions are Gaussian. Since physical artifacts induce heterogeneous distortions, bounding the effective disturbance energy encourages a corruption-agnostic, task-consistent feature prior without modeling the corruption distribution. Variable schedules introduce non-stationary targets that impair convergence, making fixed σ_f outperform linear or cosine schedules under the same budget. Thus, fixed high magnitude noise induces a stationary objective that promotes feature invariance over specific overfitting. Moreover, high variance Gaussian training smooths the denoiser Jacobian, favoring a Lipschitz bounded mapping and reducing sensitivity to unseen distortions. This fixed high noise training optimizes a robustness surrogate by enforcing BEV invariance within a bounded perturbation.

Lightweight self-supervised denoiser. As illustrated in Fig. 2b, we inject Gaussian noise $\varepsilon_{i,s} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ into the multi-scale BEV features $\mathbf{F}_{i,s}$ and train a lightweight denoiser $f_{\text{denoise}}(\cdot)$ to reconstruct the original, task-relevant representations, treating the unperturbed intermediate feature as a self-supervised proxy target:

$$\hat{\mathbf{F}}_{i,s} = f_{\text{denoise}}(\mathbf{F}_{i,s} + \varepsilon_{i,s}). \tag{7}$$

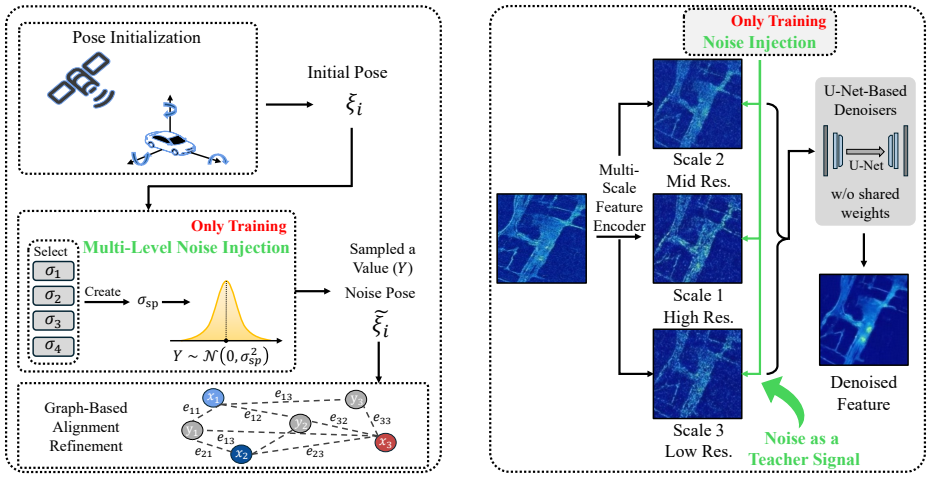
We use a self-supervised consistency loss to encourage feature stability under Gaussian perturbations across multiple scales:

$$\mathcal{L}_{\text{denoise}} = \frac{1}{S}\sum_{s=1}^{S} \|\hat{\mathbf{F}}_{i,s} - \mathbf{F}_{i,s}\|_2^2. \tag{8}$$

The denoiser employs a lightweight U-Net architecture with skip connections. This compact design is computationally efficient and introduces minimal overhead.

Relation to Denoising Autoencoders (DAEs). Although FDM adopts a noise-and-denoise form reminiscent of DAEs [31,32], its goal and setting are fundamentally different. Canonical DAEs typically pre-train on static inputs to learn a data manifold via reconstruction. In contrast, FDM serves as an in-network regularizer operating on dynamically updated intermediate BEV features and is jointly optimized with the downstream task. By bounding the disturbance energy with a fixed high-variance noise $\sigma^\star$, it optimizes a cross-noise robust objective.

$$\mathcal{L}_{\text{robust}}(\theta; \sigma^\star) = \sup_{0 \le \tau \le \sigma^\star} \mathbb{E}_{f,\varepsilon_\tau}\big[\, \|D_\theta(f + \varepsilon_\tau) - f\|_2^2 \,\big]. \tag{9}$$



(a) Noise injection in PCM. (b) Noise injection in FDM.

Fig. 2: Noise injection in PCM and FDM (training only). Left: Perturb poses $\boldsymbol{\xi}_i$ with multi-level noise $\sigma_{sp}{:=}(\sigma_t, \sigma_r)$, $\sigma_t, \sigma_r \in S$, to obtain $\tilde{\boldsymbol{\xi}}_i$, applying graph-based alignment refinement to obtain calibrated poses $\hat{\boldsymbol{\xi}}_i$. **Right:** Inject noise into multi-scale BEV features and train U-Net denoisers to recover noise-invariant representations.

This enforces task-consistent stability across a continuum of weaker, heterogeneous real-world corruptions, rather than fitting a specific noise distribution.

Joint optimization and avoiding identity mapping. The full perception framework is trained jointly with the downstream detection task:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{denoise}}, \tag{10}$$

where λ is fixed to 1.0. While feature matching could in principle lead to an identity mapping, our target $\mathbf{F}_{i,s}$ is dynamically shaped by $\mathcal{L}_{det}$ rather than being fixed. Because the injected synthetic noise σ_f is designed to be significantly larger than the inherent sensor perturbations, together with the U-Net bottleneck, the denoiser is constrained to recover task-consistent semantic structures that improve box prediction, instead of overfitting to the inherent low-level sensor artifacts present in the proxy target, as shown in Tab. 6.

4 Experiments

We evaluate **NoiseGT** on collaborative 3D object detection benchmarks to assess robustness under pose and feature corruptions. We report 3D detection Average Precision (AP) at Intersection-over-Union (IoU) thresholds 0.5 and 0.7.

4.1 Experimental Setup

Datasets. We evaluate our method on **V2X-Sim 2.0** [18], **OPV2V** [37], and **DAIR-V2X** [40]. These benchmarks include both CARLA-based [6] simulated

Table 1: Performance under pose noise. AP@0.5/AP@0.7 with pose noise (σ_t/σ_r: translation/rotation std in m/deg). Best in **bold**, second-best underlined. **NoiseGT** shows consistently smaller performance drops as σ_t/σ_r increases.

	V2X-Sim				DAIR-V2X				OPV2V			
Method / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
AP@0.5 ↑												
V2VNet [34]	0.838	0.771	0.672	0.602	0.686	0.648	0.619	0.597	0.957	0.930	0.885	0.849
F-Cooper [3]	0.630	0.534	0.470	0.446	0.727	0.697	0.677	0.663	0.886	0.752	0.642	0.574
V2X-ViT [36]	0.885	0.818	0.730	0.676	0.709	0.692	0.670	0.654	0.944	0.929	0.891	0.846
Where2comm [11]	0.736	0.619	0.549	0.505	0.696	0.683	0.665	0.655	0.902	0.829	0.748	0.694
CoSDH [35]	0.877	0.765	0.602	0.542	0.723	0.661	0.625	0.608	0.919	0.860	0.758	0.702
CoAlign [21]	0.882	0.830	0.750	0.665	0.745	0.724	0.692	0.673	0.967	0.948	0.904	0.849
Ours (NoiseGT)	**0.896**	**0.859**	**0.799**	**0.745**	**0.747**	**0.731**	**0.704**	**0.691**	**0.968**	**0.958**	**0.931**	**0.901**
AP@0.7 ↑												
V2VNet [34]	0.633	0.506	0.434	0.394	0.434	0.408	0.387	0.374	0.863	0.797	0.723	0.683
F-Cooper [3]	0.484	0.381	0.361	0.357	0.544	0.530	0.519	0.509	0.702	0.513	0.474	0.451
V2X-ViT [36]	0.730	0.674	0.621	0.588	0.546	0.540	0.533	0.526	0.843	0.828	0.792	0.757
Where2comm [11]	0.575	0.418	0.401	0.399	0.530	0.524	0.516	0.510	0.693	0.614	0.563	0.533
CoSDH [35]	0.821	0.559	0.445	0.432	0.464	0.422	0.411	0.408	0.853	0.667	0.564	0.539
CoAlign [21]	0.838	0.714	0.638	0.577	**0.594**	0.579	0.565	0.559	**0.919**	0.864	0.801	0.733
Ours (NoiseGT)	**0.849**	**0.754**	**0.710**	**0.661**	0.593	**0.581**	**0.568**	**0.564**	**0.919**	**0.887**	**0.844**	**0.804**

and real-world cooperative perception scenarios, with approximately 9k–12k frames and varying sensing ranges.

Implementation details. We build NoiseGT on OpenCOOD [37], using the PointPillars [16] encoder. To verify generalizability, we also evaluate SECOND [38] and VoxelNet [42] under identical settings. For denoising, we apply lightweight U-Net modules to multi-scale backbone features, setting the weight of $\mathcal{L}_{\text{denoise}}$ to 1.0. Smaller values under-regularize and fail under strong noise, while larger ones over-regularize and reduce gains. To further stress-test generalization, we train PCM with pose noise sampled from $\mathcal{S} = \{0.0, 0.2, 0.4, 0.6\}$, while evaluating on more severe noise levels up to 1.2 to assess robustness to unseen magnitudes. We evaluate not only i.i.d. perturbations but also Laplace and structured BEV corruptions to better approximate real-world feature missingness.

4.2 Quantitative Results

We evaluate three datasets under pose noise with translation σ_t (m) and rotation σ_r (°). Tab. 1 reports four noise levels from $(0.0\,\text{m}, 0.0°)$ to $(1.2\,\text{m}, 1.2°)$. While all baselines degrade with increasing noise, NoiseGT degrades slowest, indicating robustness. On DAIR-V2X, denser scenes make association and refinement harder, which reduces the gains. NoiseGT matches the clean performance and improves under noise. Unless otherwise specified, all percentages reported in this paper denote relative improvements over the corresponding baseline.

Robustness to realistic corruptions. Real-world collaborative perception suffers from heterogeneous drift, correlated jitter, outliers, and structured

Table 2: Robustness under non-i.i.d. pose noise. NoiseGT consistently outperforms CoAlign under heterogeneous, correlated jitter and outlier perturbations.

Method / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
AP@0.5 ↑				
CoAlign [21]	0.739	0.706	0.683	0.670
Ours (NoiseGT)	**0.742**	**0.718**	**0.698**	**0.687**
AP@0.7 ↑				
CoAlign [21]	0.587	0.570	0.561	0.555
Ours (NoiseGT)	**0.588**	**0.573**	**0.566**	**0.561**

Table 3: Robustness under coupled perturbations. With BEV grid dropout enabled at inference, NoiseGT maintains higher accuracy across all pose noise levels.

Method / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
AP@0.5 ↑				
CoAlign [21]	0.610	0.586	0.549	0.525
Ours (NoiseGT)	**0.636**	**0.609**	**0.582**	**0.564**
AP@0.7 ↑				
CoAlign [21]	0.452	0.431	0.411	0.397
Ours (NoiseGT)	**0.464**	**0.450**	**0.433**	**0.428**

feature missingness. To approximate these conditions, we evaluate NoiseGT on DAIR-V2X under two stress tests: **(i) Non-i.i.d. pose noise.** We simulate real-world drift and outliers by assigning 30% of agents as "noisy" (2.5× scale), adding frame-level correlated jitter (σ_t=0.15 m, σ_r=0.5°), and injecting rare outliers (15% probability). NoiseGT consistently outperforms CoAlign [21] (Tab. 2), demonstrating superior robustness to heterogeneous and correlated noise. **(ii) Coupled perturbations.** To model coupled localization errors and structured feature missingness, we apply DropBlock-style BEV grid dropout during inference across varying pose noise levels. The dropout is applied to multi-scale BEV features before fusion and follows a standard block masking rule with p=0.2 and base block size b_0=8. Despite training solely with Gaussian noise, NoiseGT generalizes effectively to this structured missingness. This behavior is theoretically supported by a Lipschitz-style stability bound: DropBlock-style masking can be rewritten as an unbiased structured perturbation with controlled energy, and thus the output deviation is bounded by the local Lipschitz constant of the feature-to-prediction mapping. The widening performance gap over CoAlign [21] under severe noise (Tab. 3) suggests that our fixed perturbation strategy encourages a corruption-agnostic representation.

Computational efficiency and scalability. During inference, NoiseGT disables noise injection and runs a single end-to-end inference pipeline, whose latency and memory footprint include the single-agent backbone feature extraction, the full PCM pose-graph optimization, and the complete detection forward pass with FDM enabled. As shown in Tab. 4, it is substantially more latency- and memory-efficient than iterative diffusion methods like CoDiff [14]. Under a compute-matched single-step setting, NoiseGT consistently achieves higher AP@0.5 and AP@0.7 across all noise levels, yielding a superior accuracy–cost trade-off. Regarding scalability, the PCM pose graph is naturally bounded in real V2X deployments: physical bandwidth and sensing ranges limit participating agents and co-observed objects, capping the optimization size and preventing computational blow-up even in dense traffic.

Table 4: Budget-matched comparison with CoDiff on DAIR-V2X. We report inference latency, GPU memory, and detection accuracy (AP@0.5 / AP@0.7) under different pose noise levels. For budget matching, CoDiff uses one diffusion step (N=1), so the reported latency is for N=1, in general $T(N) \approx 127.7 \times N$ ms. Under the same budget, NoiseGT achieves higher accuracy with substantially lower cost.

	Efficiency ↓		Accuracy (AP@0.5 / AP@0.7) ↑			
Method / Metric	Time (ms)	Mem (MB)	0.0 / 0.0	0.4 / 0.4	0.8 / 0.8	1.2 / 1.2
CoDiff [14]	127.7	769.17	0.716 / 0.504	0.682 / 0.479	0.660 / 0.473	0.652 / 0.471
Ours (NoiseGT)	**60.03**	**341.09**	**0.747** / **0.593**	**0.731** / **0.581**	**0.704** / **0.568**	**0.691** / **0.564**

4.3 Ablation and Analysis

We conduct ablation studies on V2X-Sim to quantify the contribution of each component, as summarized in Tab. 7. Module A denotes **multi-level noise injection** in PCM, while Module B represents the **feature denoiser** in FDM. The baseline removes both Module A and Module B for reference.

Effect of Module A. Module A slightly reduces performance under clean noise (0.824 vs. 0.838) but becomes increasingly beneficial as pose noise grows, reaching 6.1% at $(\sigma_t, \sigma_r) = (1.2\,\text{m}, 1.2^\circ)$. This indicates that multi-level noise sampling improves robustness across perturbation scales.

Effect of Module B. Module B improves performance under clean noise (0.851 vs. 0.838 at $(0.0\,\text{m}, 0.0^\circ)$), supporting the benefit of self-supervised feature denoising, though gains diminish under severe pose noise.

Combined effect of Modules A and B. Combining Module A and Module B yields the best results, reaching 0.661 under the strongest noise $(1.2\,\text{m}, 1.2^\circ)$ versus 0.577 for the baseline. The two modules are complementary: Module A mitigates pose uncertainty via PCM, while Module B reduces feature corruption via FDM. Module A alone can slightly hurt performance, because bbox-based pose-graph refinement may introduce a small micro-drift even under clean initialization. This behavior is not specific to our method and is also observed in the CoAlign baseline. Adding Module B restores and further improves clean performance to 0.849, confirming the benefit of both components and the strongest effect when used together.

Diagnostic under coupled noise. In real-world scenarios, spatial misalignment and feature corruption are inherently coupled. Our diagnostic tests on V2X-Sim reveal that under coupled noise, the baseline AP@0.7 degrades severely to 0.5211. While PCM primarily mitigates pose errors and FDM targets feature corruption, relying on either module alone is insufficient. Joint mitigation is important to handle compounded failure modes, improving the coupled performance to 0.5634.

Effect of explicit denoising. To show that NoiseGT's gains on V2X-Sim are not solely from noise augmentation and that FDM does not collapse to an identity mapping, we compare four variants in Tab. 5. These are **Off**, which removes FDM, **Gaussian**, which applies fixed smoothing, **Arch-only**,

Table 5: Baselines for FDM on V2X-Sim. We compare removing FDM (Off), fixed Gaussian smoothing, a parameter-matched learnable refinement baseline (Arch-only), and NoiseGT. Best in **bold**, second-best underlined.

	AP@0.5 ↑				AP@0.7 ↑			
Method / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
Off (no FDM)	0.868	0.834	0.766	0.717	0.828	0.730	0.675	0.634
Gaussian smoothing	0.866	0.828	0.767	0.707	0.811	0.713	0.659	0.613
Arch-only (param-matched)	0.883	0.844	0.778	0.726	0.841	0.740	0.690	0.654
Ours (NoiseGT)	**0.896**	**0.859**	**0.799**	**0.745**	**0.849**	**0.754**	**0.710**	**0.661**

a parameter-matched learnable refinement trained without feature noise injection or $\mathcal{L}_{\text{denoise}}$, and **Ours** (NoiseGT). Arch-only consistently improves over Off, confirming a capacity gain from adding a learnable refinement block. NoiseGT further improves over Arch-only across all pose-noise levels, demonstrating additional gains from explicit noise-injection and denoising learning beyond architecture alone. Gaussian performs worst and blurs activations, whereas NoiseGT preserves sharper task-relevant boundaries.

Table 6: Effectiveness of explicit denoising. NoiseGT outperforms the training-only noise augmentation baseline across pose noise levels, indicating gains from explicit denoising learning rather than augmentation.

	AP@0.5 ↑				AP@0.7 ↑			
Method / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
Noise aug-only	0.874	0.838	0.769	0.714	0.797	0.708	0.652	0.612
Ours (NoiseGT)	**0.896**	**0.859**	**0.799**	**0.745**	**0.849**	**0.754**	**0.710**	**0.661**

Feature noise magnitude analysis. We further evaluate different feature noise levels $\sigma_f \in \{0.05, 0.1, 0.2, 0.3\}$ across varying pose noises. As shown in Fig. 3, a moderate level (σ_f=0.2) yields the best performance, especially under stronger pose noise. Too little noise causes overfitting to minor perturbations, while excessive noise slightly harms clean performance. These results confirm that injecting sufficiently strong and consistent noise helps the model learn a stable denoising prior and enhances robustness to real-world feature corruptions.

Solver generalization. To test whether PCM is solver-agnostic, we replace the default solver with RoCo [13]. As shown in Tab. 8, multi-level noise injection consistently improves RoCo across all noise levels. Under severe noise $(1.2\,\text{m}, 1.2°)$, our framework achieves relative gains of 4.27% and 4.41% in AP@0.5 and AP@0.7, and it also yields 3.52% and 5.14% relative gains under clean noise $(0.0\,\text{m}, 0.0°)$. These results indicate that our noise-driven paradigm is a universal, plug-and-play enhancer for alignment pipelines.

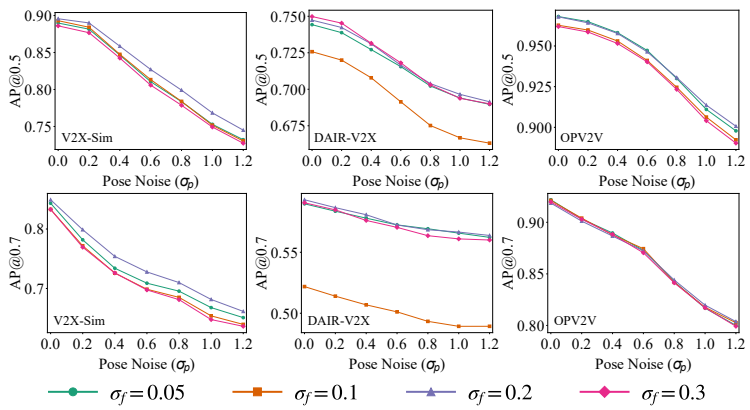


Fig. 3: Impact of feature noise level σ_f under varying pose noise magnitudes. Models trained with different $\sigma_f \in \{0.05, 0.1, 0.2, 0.3\}$ are evaluated under increasing pose noise. A moderate noise level (σ_f=0.2) yields the most stable and robust performance, confirming the effectiveness of our **high-noise-over-low-noise** strategy.

Robustness to residual pose errors. We further evaluate the pose refinement behavior on DAIR-V2X using the absolute median pairwise relative pose error (RPE). Since our PCM is solver-agnostic and utilizes the identical off-the-shelf pose-graph optimizer as CoAlign [21] during inference, both methods naturally yield identical absolute RPEs after refinement. For example, under the most severe noise setting, both methods are left with a substantial residual median translation error of 2.15 m. However, despite confronting the exact same uncorrected residual misalignments, NoiseGT achieves significantly higher downstream detection accuracy. This compellingly confirms that our multi-level noise injection during training effectively equips the downstream network with distributionally robust tolerance to inherent pose errors, rather than merely relying on the physical pose solver.

Backbone generalization. To verify that NoiseGT is detector-agnostic, we replace PointPillars with **SECOND** [38] and **VoxelNet** [42] on V2X-Sim while keeping all other protocols unchanged. As shown in Tab. 9, NoiseGT consistently outperforms CoAlign [21] under both backbones across pose noise levels for AP@0.5/0.7. The gains increase under moderate and severe noise, indicating that our noise-driven calibration and feature refinement complement different backbones and generalize beyond a specific BEV encoder.

4.4 Qualitative Results

As shown in Fig. 4, we present qualitative detection results on the **V2X-Sim** dataset under severe localization perturbation (1.2 m, 1.2°). Compared with representative collaborative perception methods (Where2comm [11], CoAlign [21], and CoSDH [35]), our approach produces noticeably cleaner and more spatially consistent bounding boxes. The baseline methods exhibit evident misalignments

Table 7: Ablation study. Module A is **multi-level noise injection**, and module B is the **feature denoiser**.

Setting / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
	AP@0.7 ↑			
Baseline (w/o A,B)	0.838	0.714	0.638	0.577
Module A only	0.824	0.735	0.681	0.638
Module B only	0.851	0.706	0.645	0.577
Module A+B	0.849	0.754	0.710	0.661

Table 8: Solver generalization. We integrate NoiseGT with RoCo [13] to evaluate its plug-and-play capability.

Solver / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
	AP@0.5 ↑			
RoCo (Original)	0.853	0.803	0.746	0.703
Ours (w/ RoCo)	**0.883**	**0.838**	**0.778**	**0.733**
	AP@0.7 ↑			
RoCo (Original)	0.798	0.682	0.640	0.612
Ours (w/ RoCo)	**0.839**	**0.715**	**0.674**	**0.639**

Table 9: Backbone generalization. We evaluate our method using **SECOND** and **VoxelNet** backbones on V2X-Sim under various pose noise levels.

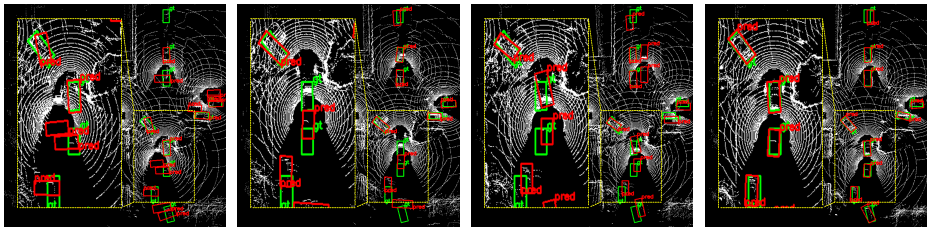
	AP@0.5 ↑				AP@0.7 ↑			
Backbone / Metric	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2	0.0/0.0	0.4/0.4	0.8/0.8	1.2/1.2
CoAlign (SECOND)	0.888	0.834	0.732	0.667	**0.790**	0.645	0.563	0.511
Ours (SECOND)	**0.889**	**0.841**	**0.777**	**0.720**	0.771	**0.666**	**0.609**	**0.560**
CoAlign (VoxelNet)	0.888	0.841	0.755	0.683	0.840	0.710	0.652	0.592
Ours (VoxelNet)	**0.898**	**0.859**	**0.795**	**0.741**	**0.853**	**0.759**	**0.716**	**0.667**

and missed detections due to inaccurate pose fusion, whereas our framework preserves precise localization and compact bounding boxes that align well with the underlying point clouds. These visual results further corroborate the strong robustness of NoiseGT against severe spatial misalignments.

5 Conclusion

In this paper, we presented **NoiseGT**, a noise-driven framework for robust collaborative perception under coupled pose and feature uncertainties. By integrating the Pose Calibration Module (PCM) and the Feature Denoising Module (FDM), NoiseGT leverages noise as an empirical self-supervised teacher to promote cross-noise-invariant representations. Extensive experiments on V2X-Sim, DAIR-V2X, and OPV2V benchmarks demonstrate state-of-the-art performance under severe perturbations. Rather than attempting to reconstruct exact noise-free targets, our paradigm establishes that consistent noise injection can serve as an effective regularization signal, learning task-relevant stable features while significantly reducing the reliance on clean intermediate supervision. Furthermore, its lightweight, diffusion-free design ensures strong computational efficiency, making it highly scalable for real-time multi-agent deployments.

Limitations and Future Work. Despite its effectiveness, NoiseGT has practical limitations. First, extreme pose errors (*e.g.*, up to $6.0\,\text{m}, 6.0^\circ$) remain a bottleneck for current bounding-box-based graph optimization. Second, under



(a) Where2comm **(b)** CoAlign **(c)** CoSDH **(d)** NoiseGT

Fig. 4: Qualitative comparison on V2X-Sim under severe localization noise $(\sigma_t, \sigma_r) = (1.2\,\text{m}, 1.2^\circ)$. **NoiseGT** produces cleaner and more spatially consistent bounding boxes than **Where2comm**, **CoAlign**, and **CoSDH**, demonstrating strong robustness to both pose and feature noise.

perfectly noise-free initializations, inherent observation noise from box predictions can induce a slight micro-drift during pose refinement. Lastly, non-zero-mean feature bias may still require more adaptive uncertainty modeling. Future work will explore confidence-aware, adaptive regularization to dynamically adjust refinement strength across varied clean and corrupted environments, further advancing the reliability of real-world autonomous driving.

Acknowledgements

This work is supported by the NSFC under Grant 62472369.

References

1. Arnold, E., Mozaffari, S., Dianati, M.: Fast and robust registration of partially overlapping point clouds. IEEE Robotics and Automation Letters **7**(2), 1502–1509 (2021)
2. Chen, H., Wu, W., Zhang, S., Wu, C., Zhong, R.: A gnss/lidar/imu pose estimation system based on collaborative fusion of factor map and filtering. Remote Sensing **15**(3), 790 (2023)
3. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. pp. 88–100 (2019)
4. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 514–524. IEEE (2019)
5. Chiu, H.k., Hachiuma, R., Wang, C.Y., Smith, S.F., Wang, Y.C.F., Chen, M.H.: V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. arXiv preprint arXiv:2502.09980 (2025)
6. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning. pp. 1–16 (2017)

7. Feng, D., Wang, Z., Zhou, Y., Rosenbaum, L., Timm, F., Dietmayer, K., Tomizuka, M., Zhan, W.: Labels are not perfect: Inferring spatial uncertainty in object detection. IEEE Transactions on Intelligent Transportation Systems **23**(8), 9981–9994 (2021)
8. Hao, J., Sun, L., Xiang, T., Wan, Y., Song, H., Lv, P.: Feakm: Robust collaborative perception under noisy pose conditions. In: Proceedings of the 2024 4th International Joint Conference on Robotics and Artificial Intelligence. pp. 98–102 (2024)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Hong, S., Liu, Y., Li, Z., Li, S., He, Y.: Multi-agent collaborative perception via motion-aware robust communication network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15301–15310 (2024)
11. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
12. Huang, X., Wang, J., Xia, Q., Chen, S., Yang, B., Li, X., Wang, C., Wen, C.: V2x-r: Cooperative lidar-4d radar fusion with denoising diffusion for 3d object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27390–27400 (2025)
13. Huang, Z., Wang, S., Wang, Y., Li, W., Li, D., Wang, L.: Roco: Robust cooperative perception by iterative object matching and pose adjustment. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7833–7842 (2024)
14. Huang, Z., Wang, S., Wang, Y., Wang, L.: Codiff: Conditional diffusion model for collaborative 3d object detection (2025), https://arxiv.org/abs/2502.14891
15. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4037–4058 (2020)
16. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
17. Li, C., Xu, L., Jin, C., Wang, L.: Graphps: Graph pair sequences-based noisy-robust multi-hop collaborative perception. IEEE Transactions on Intelligent Vehicles **9**(2), 3895–3905 (2023)
18. Li, Y., Ma, D., An, Z., Wang, Z., Zhong, Y., Chen, S., Feng, C.: V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. IEEE Robotics and Automation Letters **7**(4), 10914–10921 (2022). https://doi.org/10.1109/LRA.2022.3192802
19. Liu, Y.C., Tian, J., Glaser, N., Kira, Z.: When2com: Multi-agent perception via communication graph grouping. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4106–4115 (2020)
20. Liu, Y.C., Tian, J., Ma, C.Y., Glaser, N., Kuo, C.W., Kira, Z.: Who2com: Collaborative perception via learnable handshake communication. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 6876–6883. IEEE (2020)
21. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4812–4818. IEEE (2023)
22. Mo, Y., Vijay, R., Rufus, R., Boer, N.d., Kim, J., Yu, M.: Enhanced perception for autonomous vehicles at obstructed intersections: An implementation of vehicle to infrastructure (v2i) collaboration. Sensors **24**(3), 936 (2024)

23. Nachkov, A., Paudel, D.P., Danelljan, M., Van Gool, L.: Diffusion-based particle-detr for bev perception. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2725–2735. IEEE (2025)
24. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749 (2021)
25. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001)
26. Shi, S., Cui, J., Jiang, Z., Yan, Z., Xing, G., Niu, J., Ouyang, Z.: Vips: Real-time perception fusion for infrastructure-assisted autonomous driving. In: Proceedings of the 28th annual international conference on mobile computing and networking. pp. 133–146 (2022)
27. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
28. Song, Z., Wen, F., Zhang, H., Li, J.: A cooperative perception system robust to localization errors. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–6. IEEE (2023)
29. Song, Z., Xie, T., Zhang, H., Liu, J., Wen, F., Li, J.: A spatial calibration method for robust cooperative perception. IEEE Robotics and Automation Letters **9**(5), 4011–4018 (2024)
30. Song, Z., Yang, L., Xu, S., Liu, L., Xu, D., Jia, C., Jia, F., Wang, L.: Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection. In: European Conference on Computer Vision. pp. 347–366. Springer (2024)
31. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)
32. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(12) (2010)
33. Wang, J., Nordström, T.: Latency robust cooperative perception using asynchronous feature fusion. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1–10. IEEE (2025)
34. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: European conference on computer vision. pp. 605–621. Springer (2020)
35. Xu, J., Zhang, Y., Cai, Z., Huang, D.: Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 6834–6843 (June 2025)
36. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022)
37. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)
38. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)

39. Ye, X., Yaman, B., Cheng, S., Tao, F., Mallik, A., Ren, L.: Bevdiffuser: Plug-and-play diffusion model for bev denoising with ground-truth guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1495–1504 (2025)
40. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., Nie, Z.: 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 (CVPR). pp. 21361–21370 (June 2022)
41. Yuan, Y., Cheng, H., Sester, M.: Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. IEEE Robotics and Automation Letters **7**(2), 3054–3061 (2022)
42. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)
43. Zou, J., Tian, K., Zhu, Z., Ye, Y., Wang, X.: Diffbev: Conditional diffusion model for bird's eye view perception. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7846–7854 (2024)