RECO: Region-Aware Compensation for Extrinsic Perturbations in Roadside 3D Detection

Junsheng Du[1], Zhaocheng He[1(✉)], and Yuhuan Lu[2(✉)]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China
dujsh3@mail2.sysu.edu.cn, hezhch@mail.sysu.edu.cn
[2] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China
yhlu@mpu.edu.mo

Abstract. In intelligent transportation systems, roadside 3D object detection provides wide-area perception crucial for traffic understanding, cooperative early warning, and safe autonomous driving. However, existing methods suffer from high sensitivity to camera extrinsics; even slight deviations (whether manifesting as transient jitter or persistent drift) can be significantly amplified by projective geometry. This cascade results in severe feature misalignment and degraded localization. To mitigate this limitation, we propose RECO, a region-aware extrinsic compensation framework that corrects extrinsics using piecewise 6-DoF pose offsets. RECO predicts a learnable range boundary to partition the scene into near and far regions, estimating region-specific pose corrections. A differentiable sigmoid gate then smoothly blends the two compensated geometries to preserve continuous BEV sampling and facilitate stable optimization. To supervise the refinement of extrinsics, we introduce an auxiliary reprojection loss that compares 2D bounding boxes projected from 3D ground truth against 2D annotations, optimizing it jointly with the standard detection objective. Extensive experiments on the DAIR-V2X-I and Rope3D benchmarks under extrinsic perturbations demonstrate consistent improvements over state-of-the-art baselines across both yaw and z-axis deviations. RECO also generalizes from transient perturbations to persistent shifts, maintaining highly competitive performance under strict calibration uncertainty.

Keywords: Vehicle-to-Infrastructure · 3D detection · Camera extrinsics

1 Introduction

Roadside cameras provide a global viewpoint that covers intersections and long road segments well beyond the sensing horizon of on-board sensors, forming the backbone of applications such as early warning and autonomous driving systems [9, 22]. However, vision-based roadside 3D object detection critically depends on accurate and static camera extrinsics. Small pose perturbations can be magnified by the geometric projection to the ground plane, leading to severe

✉ Corresponding author.

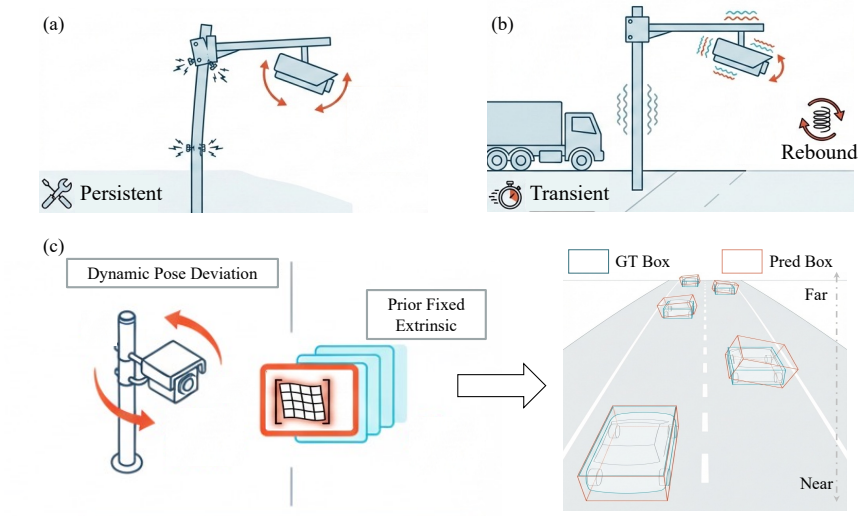

Fig. 1: Illustration of camera pose deviations and their impact on roadside 3D detection. (a) Persistent pose drift caused by long-term structural changes. (b) Transient pose perturbations induced by external disturbances (e.g., vibrations) that may rebound to the nominal pose. (c) Using a fixed, pre-calibrated extrinsic pose under drift causes misaligned 3D predictions, with errors typically scaling proportionally with object distance.

feature and spatial misalignment, a phenomenon that drastically worsens at long range.

Recent works [23, 25, 28, 29] have improved robustness to camera pose jitter by mitigating sensitivity to depth estimation and stabilizing the bird's-eye-vision (BEV) transformation. These approaches typically introduce more robust BEV construction designs for image-to-BEV lifting, such as incorporating height cues to reduce brittleness under calibration noise [29]. Meanwhile, other architectures focus on making BEV aggregation smoother under calibration noise by enforcing stronger spatial alignment during fusion [25]. Collectively, these advances suggest that perception robustness under imperfect calibration benefits from jointly optimizing geometric lifting, BEV pooling, and alignment-aware fusion.

However, the aforementioned methods typically assume that camera extrinsics are fixed parameters. In real-world deployments, roadside camera poses are rarely static: thermal expansion, structural deformation, and sensor remounting can introduce subtle yet persistent drifts (Fig. 1(a)). Moreover, transient perturbations may occur under acute external disturbances (e.g., vibrations induced by heavy commercial vehicles) and subsequently rebound to the nominal pose shortly after (Fig. 1(b)). Under such dynamic pose variations, projection relying on a stale prior extrinsic inevitably induces spatial misalignment and degrades overall 3D detection performance (Fig. 1(c)).

Conversely, calibration-free methods aim to enable roadside perception without explicit extrinsic priors. Building on this, representative approaches achieve BEV perception by learning implicit image-to-BEV transformations or incorporating geometric cues provided by V2X infrastructure [2, 33]. While effectively reducing reliance on rigorous extrinsic calibration, these approaches often compromise geometric consistency and introduce dependencies on anchor accuracy or communication reliability, which can hinder deployment scalability.

To bridge this gap and explicitly handle varying extrinsic drift, we propose RECO, a **RE**gion-based **CO**mpensation mechanism that models extrinsic refinement as piecewise 6-DoF pose offsets spanning rotation (roll, pitch, yaw) and translation (t_x, t_y, t_z). RECO predicts a spatial range boundary to partition the scene into near and far regions, estimating two SE(3) corrections to refine the nominal extrinsics accordingly. To prevent spatial discontinuities induced by hard spatial assignments, we introduce a differentiable soft gate that smoothly blends the two extrinsic compensations, yielding a continuous projection field and well-behaved gradients for end-to-end training. This design yields stable BEV feature sampling despite transient camera jitter and mitigates abrupt feature artifacts caused by pose variations.

We evaluate RECO on two primary roadside 3D detection benchmarks, DAIR-V2X-I [32], and Rope3D [30], comparing it against strong BEV-based baselines. To systematically quantify robustness under controlled perturbations, we train and evaluate the models with transient extrinsic noise modeled by a zero-mean Gaussian distribution. Furthermore, we assess cross-condition generalization by applying mean-shifted extrinsic offsets during inference. Across both datasets, RECO consistently elevates detection robustness under transient perturbations and maintains strong performance when evaluated under persistent drift. In addition, ablation studies confirm that introducing the proposed compensation mechanism, particularly the projection with piecewise 6-DoF pose offsets, accounts for the majority of robustness gains and effectively neutralizes projection misalignment. In summary, our contributions are as follows:

- **Region-aware 6-DoF extrinsic compensation**: RECO predicts a range boundary and near/far 6-DoF pose offsets to refine nominal camera extrinsics in a spatially-adaptive, piecewise manner.
- **Differentiable soft gating**: We introduce a sigmoid-based soft gate to smoothly interpolate near and far compensated projections, guaranteeing mathematically continuous feature fields and stable end-to-end optimization.
- **Strong empirical results**: RECO achieves state-of-the-art robustness on DAIR-V2X-I and Rope3D, particularly under rigorously simulated extrinsic perturbations.

2 Related Work

We review relevant works on roadside monocular BEV-based 3D detection and 6-DoF/SE(3) extrinsic estimation, covering BEV-based designs for infrastructure perception and learning-based calibration methods.

Roadside monocular 3D object detection. Most BEV-based 3D detectors construct an image-to-BEV mapping using calibrated camera extrinsics and treat them as fixed inputs, including voxel pipelines and transformer-style BEV formulations [5, 6, 11, 12, 16]. For roadside monocular perception, recent works further specialize BEV construction to mitigate depth ambiguity and instability caused by discretization, improving robustness under calibration noise while still relying on nominal extrinsics [23, 25, 28, 29]. In parallel, calibration-free approaches explore implicit image-to-BEV transformations or external geometric cues (e.g., segmentation, V2X positions) to reduce reliance on explicit calibration [2, 33]. Recent roadside monocular designs also leverage stronger priors (e.g., 2D detection prompts) to improve BEV localization and long-range stability, yet they typically do not explicitly model range-wise extrinsic drift within the projection operator [15]. Overall, existing works largely improve robustness by stabilizing BEV lifting/pooling/alignment or by bypassing calibration, leaving room for approaches that adapt projection geometry online while preserving metric consistency under deployment-time extrinsic drift.

6-DoF / SE(3) estimation. A widely used strategy for handling calibration drift is to estimate a 6-DoF offset or SE(3) correction that refines nominal extrinsics from cross-modal alignment cues, instead of assuming fixed calibration. Early self-calibration networks regress SE(3) directly from projected depth/RGB feature or jointly enforce geometric consistency, enabling online correction under miscalibration [8, 14, 19]. More recent designs improve robustness by explicitly modeling cross-modal correspondence with cost volumes, correlation layers, or transformer-style association, followed by end-to-end regression of SE(3) updates [27]. Related to extrinsic refinement, image-to-lidar registration treats SE(3) as a relative pose between modalities and learns robust pose solvers for challenging cross-domain alignment [10, 18]. In parallel, a large body of work studies SE(3) estimation for rigid point clouds, proposing learned correspondences and robust matching methods that are often adapted or reused for calibration refinement pipelines [1, 3, 7, 17, 24, 31]. Finally, target-free calibration methods optimize SE(3) by maximizing cross-modal consistency (e.g., mutual-information objectives), offering an alternative when supervised training signals are limited.

Despite these advances, existing BEV detectors largely treat extrinsics as fixed priors, and most SE(3) calibration methods are not integrated to handle range-dependent drift within BEV projection, motivating an extrinsic-aware, range-adaptive compensation mechanism with continuous projection for robust roadside detection.

3 Methodology

3.1 Problem Formulation

Given a single roadside RGB image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ and its camera extrinsics parameterized by rotation $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and translation $t \in \mathbb{R}^3$, we denote the

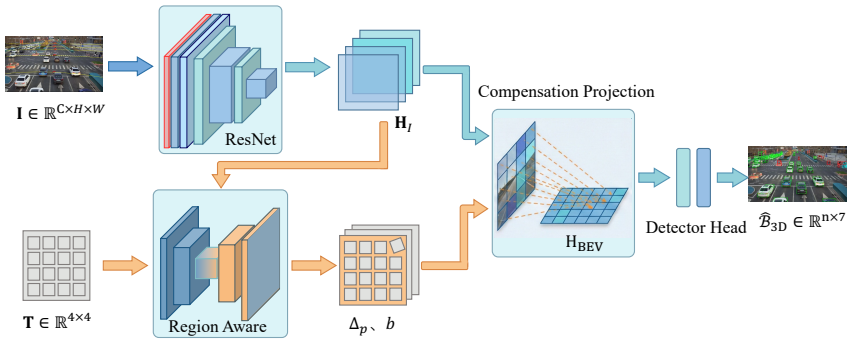

Fig. 2: Overview of RECO, a region-aware extrinsic compensation framework for infrastructure monocular 3D object detection. The region-aware module predicts near/far 6-DoF pose offsets Δ_p and a range boundary b, which are used by the compensation projection to construct BEV features for the detector head.

homogeneous transform as

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & t \\ 0^\top & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} \tag{1}$$

Our goal is to jointly estimate a set of near/far 6-DoF pose offsets $\Delta_p \in \mathbb{R}^{1\times6}$, ordered as [roll, pitch, yaw, t_x, t_y, t_z], and a spatial boundary b that partitions the scene. Using Δ_p to compensate for camera extrinsics, the model performs 3D object detection and outputs a set of 3D bounding boxes $\hat{\mathcal{B}}_{3D} \in \mathbb{R}^{N\times7}$, where each box is parameterized by its center (x, y, z), physical size (w, h, l), and heading angle θ. Formally, we seek a mapping $\mathcal{F}$ such that:

$$(\Delta_p, b, \hat{\mathcal{B}}_{3D}) = \mathcal{F}(\mathbf{I}, \mathbf{T}) \tag{2}$$

3.2 Overall Architecture

The overall pipeline is illustrated in Fig. 2. Given a monocular roadside image and its nominal camera extrinsics, we first extract multi-scale 2D features using a ResNet backbone. A region-aware extrinsic compensation network then takes the visual features alongside the prior extrinsics as input, predicting (i) 6-DoF pose offsets for near/far regions and (ii) a range boundary that partitions the scene. Using these predicted offsets, our compensation projection module lifts image features from the perspective view onto the BEV plane to construct BEV features, which are subsequently fed into a 3D detection head to produce 3D bounding box predictions.

3.3 Region-Aware Module

In the roadside BEV projection, the effect of camera extrinsic perturbations is inherently range-dependent and cannot be well captured by a single global rigid

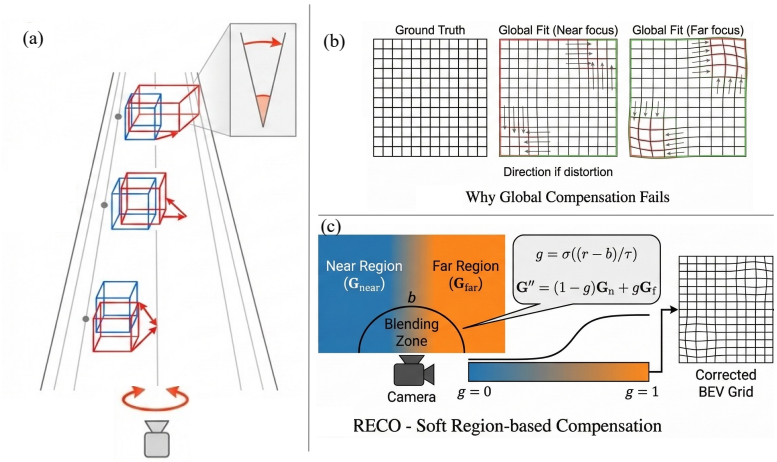

Fig. 3: Motivation and design for our region-aware extrinsic compensation. (a) Range-dependent effect of camera extrinsic perturbations: near-range misalignment is dominated by translation bias, while far-range misalignment is more sensitive to rotation jitter (the blue box denotes the ground-truth, while red denotes the prediction). (b) A single global 6-DoF correction cannot simultaneously fit near and far ranges under imbalanced supervision signals. (c) RECO predicts a range boundary and blends near/far SE(3) offsets.

correction. In particular, near-range errors are more influenced by translational bias, while far-range misalignments are typically more sensitive to rotational jitter [20], as illustrated in Fig. 3(a). Meanwhile, the supervision signals in the data (e.g., detection errors and reprojection errors) are not uniformly distributed across distances (Fig. 3(b)). Learning a single global compensation implicitly assumes uniform errors throughout the scene. In practice, optimization can be dominated by a particular distance interval, hindering the learning of a globally consistent correction.

Building on the above analysis, we implement this idea with a region-aware (RA) module that jointly predicts pose corrections and range boundaries. Given an aggregated visual feature $\mathbf{H}_I \in \mathbb{R}^{d \times H_{img} \times W_{img}}$ and nominal camera extrinsics $\mathbf{T}$, RA concatenates and encodes them into a shared latent embedding $\mathbf{Z}$. Here, d is the dimension of the image feature. To generalize near–far partitioning to k range segments, a boundary head predicts $k-1$ boundaries via a residual form:

$$b = \tilde{b} + \tanh(\mathbf{U}_b \cdot \mathbf{Z}) \tag{3}$$

where $\tilde{b}$ specifies the prior boundary [20 m, 40 m, ...], and $\mathbf{U}_b$ is a learnable weight matrix of the boundary head. In parallel, a pose head outputs 6-DoF offsets Δ_p. This design enables region-adaptive extrinsic compensation that aligns with the range-dependent error characteristics of BEV projection and downstream detection.

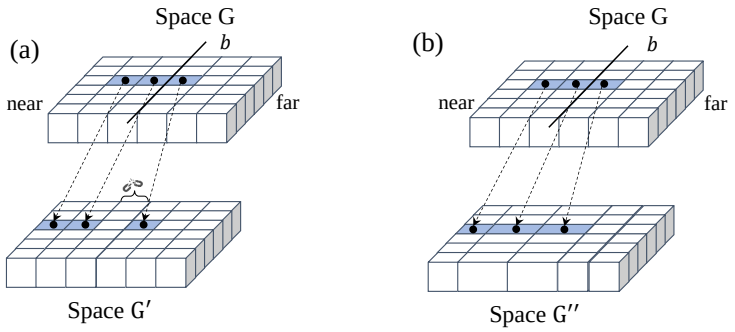

Fig. 4: Motivation for soft gating in range-aware extrinsic compensation. (a) Assigning piecewise offsets produces a piecewise projection, causing a discontinuous mapping at the boundary b (non-smooth BEV feature sampling). (b) A differentiable soft gate smoothly blends near/far geometries, yielding a continuous projection field and stable optimization. Here, $\mathbf{G}$ denotes the original projection space. $\mathbf{G}'$ and $\mathbf{G}''$ illustrate the discontinuous/continuous projection fields after different compensations.

3.4 Soft Compensation Projection Module

Using piecewise offsets can make the projection field non-smooth, introducing a discontinuity at the partition boundary (Fig. 4(a)). To address this issue, we introduce a soft compensation projection module (SCP) that enables differentiable geometry blending (Fig. 4(b)).

Range-aware geometry blending. Given the boundary b and near/far 6-DoF offsets $\{\Delta_{p_n}, \Delta_{p_f}\}$, SCP performs a range-aware projection by smoothly blending the corresponding geometries. We first convert $\{\Delta_{p_n}, \Delta_{p_f}\}$ to SE(3) corrections $\{\Delta\mathbf{T}_n, \Delta\mathbf{T}_f\}$ and left-compose each correction with the original extrinsic $\mathbf{T}$:

$$\mathbf{P}_n = \Delta\mathbf{T}_n\mathbf{T}, \qquad \mathbf{P}_f = \Delta\mathbf{T}_f\mathbf{T} \tag{4}$$

Given the corrected extrinsic $\mathbf{P}_*$, we construct the BEV sampling geometry by projecting each BEV cell center from the ego frame into the camera image. Concretely, we back-project a 3D point $\mathbf{X}_e = [x, y, z, 1]^\top$ in the ego frame, transform it to the camera frame via:

$$\mathbf{X}_c = \mathbf{P}_* \mathbf{X}_e = \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}. \tag{5}$$

Then we apply the camera intrinsics $\mathbf{K}$ to compute the pixel position (u, v) as follows:

$$\mathbf{X}_k = \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{bmatrix} = \mathbf{K} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}, \qquad u = \tilde{x}/\tilde{z}, \qquad v = \tilde{y}/\tilde{z} \tag{6}$$

Stacking (u, v) for all BEV cells forms the geometry mapping field $\mathbf{G}_*$, which is used to sample image features and construct BEV features.

Keeping other parameters fixed, the corrected transforms $\mathbf{P}_n$ and $\mathbf{P}_f$ yield a geometry mapping field $\mathbf{G}_n$ and $\mathbf{G}_f$ for BEV projection. For each BEV location (x, y), we compute its distance $r = \sqrt{x^2 + y^2}$ in the BEV plane, then via a soft gate as:

$$g = \sigma(\frac{r - b}{\tau}) \tag{7}$$

where τ controls the transition width. The final geometry field is obtained by gated interpolation:

$$\mathbf{G}'' = (1 - g)\mathbf{G}_n + g\mathbf{G}_f \tag{8}$$

which preserves a continuous projection field and avoids the non-smoothness introduced by hard near/far switching.

BEV feature construction. We construct BEV features by warping and aggregating image features onto a discrete BEV grid. Given backbone features $\mathbf{H}_I$, $\mathbf{G}''$ assigns each sample to BEV location (x, y), which is quantized into BEV indices (i, j) :

$$i = \left\lfloor \frac{x - x_{\min}}{\Delta x} \right\rfloor, \qquad j = \left\lfloor \frac{y - y_{\min}}{\Delta y} \right\rfloor \tag{9}$$

where $(x_{\min}, x_{\max}, y_{\min}, y_{\max})$ are the predefined BEV bounds, and $(\Delta x, \Delta y)$ are the resolution.

We then splat the projected image features onto the BEV grid and aggregate all contributions using scatter-based pooling. The BEV feature map $\mathbf{H}_{\text{BEV}} \in \mathbb{R}^{C \times H_{\text{bev}} \times W_{\text{bev}}}$ is computed by weighted averaging:

$$\mathbf{H}_{\text{BEV}}[:, i, j] = \frac{\sum \mathbf{U}_{bev}\, \mathbf{H}_I}{\sum \mathbf{U}_{bev} + \epsilon} \tag{10}$$

where summations are over samples assigned to cell (i, j), $\mathbf{U}_{bev}$ denotes an optional validity weight, and ϵ avoids division by zero. Then we aggregate projected image features into the BEV grid via scatter-based pooling: features assigned to the same cell are averaged and normalized by the corresponding cell count.

3D box prediction. Given the BEV feature map $\mathbf{H}_{\text{bev}}$, we adopt a standard BEV detection head to produce 3D bounding box predictions. The head first applies several BEV encoder layers to enhance spatial context, and then outputs dense regression and classification maps on the BEV grid. Following common practice in BEV-based detectors [28, 29], we decode the head outputs into a set of N boxes $\hat{\mathcal{B}}_{3D}$ in the ego frame, together with confidence scores. During training, these predictions are matched to ground truth following the assignment strategy used in the baseline, and the resulting detection loss is denoted as $\mathcal{L}_{\text{det}}$.

3.5 Training Objective

We train the detector with the standard 3D detection loss used in previous work [28, 29] and an auxiliary reprojection loss that supervises the predicted

extrinsic compensation; 3D box annotations supervise the detection loss, while 2D box annotations supervise the reprojection loss. During inference, we discard the reprojection branch and only use the predicted extrinsic compensation for geometry construction.

Specifically, given a 3D ground-truth box, we project its 3D corners to the image plane using the compensated projection to obtain a 2D box $\hat{\mathcal{B}}_{2D}$. We then penalize the discrepancy to the corresponding ground-truth 2D box $\mathcal{B}_{2D}$ with an L2 loss over valid instances:

$$\mathcal{L}_{rep} = \frac{1}{\sum m} \sum ||\hat{\mathcal{B}}_{2D} - \mathcal{B}_{2D}||_2^2 \tag{11}$$

where m denotes the valid mask. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{rep}} \mathcal{L}_{\text{rep}}, \tag{12}$$

where $\mathcal{L}_{det}$ is the 3D detection loss and λ_{rep} balances the reprojection supervision.

4 Experiments

4.1 Experiments Setup

Datasets. We evaluate RECO on two prominent benchmark datasets for roadside 3D detection: **DAIR-V2X** [32] and **Rope3D** [30]. **DAIR-V2X** is a large-scale benchmark for vehicle-infrastructure cooperative perception, providing synchronized multi-modal data captured at urban intersections using both vehicle-side and infrastructure-side sensors. **Rope3D** is a diverse dataset designed specifically to benchmark monocular 3D object detection from infrastructure cameras, featuring an extensive set of images collected across varied camera setups, viewpoints, and environmental conditions.

Implementation details. For architectural details, we apply ResNet 50 [4] as the backbone network and LSSFPN [16] to construct the 3D volumetric feature. The model is trained with AdamW [13] using a weight with decay of 10^{-4}, $\lambda_{\text{rep}} = 0.1$ and $\tau = 0.5$. The BEV grid spans [-51.2 m, 51.2 m] in width and [0 m, 140 m] in length for two datasets. For input images, we set the resolution to (864, 1536). Following [11,28], we apply image-level augmentations, including random cropping and rotation, as well as BEV-level augmentations, including random scaling, flipping, and rotation. All reproduction methods are trained for 50 epochs on 4 RTX A6000 GPUs with a batch size of 8.

4.2 Roadside 3D Detection

Comparison methods. We compare different BEV-based methods, such as BEVHeight (CVPR 2023 [29]), CoBEV (TIP 2024 [21]), BEVSpread (CVPR 2024 [23]), HeightFormer (TGRS 2024 [25]), BEVHeight++ (TPAMI 2025 [28]).

Table 1: Comparison on the DAIR-V2X-I [32] validation set under extrinsic perturbations. Following [26], We randomly sample angles and shifts from $\mathcal{N}(0, 0.5)$ for yaw and z. All methods are retrained under the same protocol for a fair comparison. We report AP for car (IoU=0.5), pedestrian (IoU = 0.25), and cyclist (IoU = 0.25). "↑" indicates that higher values are better.

Method	Deviation	$\text{Car}_{IoU=0.5}$ ↑			$\text{Ped.}_{IoU=0.25}$ ↑			$\text{Cyc.}_{IoU=0.25}$ ↑			FPS
		easy	moderate	hard	easy	moderate	hard	easy	moderate	hard	
BEVHeight	yaw	65.5595	61.7935	61.8782	10.9768	10.3565	10.4573	39.9807	38.2060	34.0681	19.81
CoBEV		31.4301	26.5978	26.9728	5.1976	5.0362	5.0865	23.3402	22.5645	20.7799	10.16
BEVSpread		67.0991	56.0408	56.1161	16.1610	15.3372	15.3543	41.4683	41.2027	33.9091	7.78
HeightFormer		38.6363	32.0780	32.1450	9.2197	8.8998	8.9994	14.9138	14.7369	18.8614	8.66
BEVHeight++		69.6915	58.2977	60.3824	5.7638	5.2098	5.2912	19.4069	18.7940	19.4149	7.80
Our		**70.4639**	**63.0142**	**63.0296**	**16.2161**	**15.3758**	**15.4261**	**43.3582**	**42.9046**	**36.4710**	17.66
BEVHeight	z	58.5738	52.8735	53.0879	3.4300	3.3005	3.3102	26.3826	25.8640	19.1211	19.82
CoBEV		26.7985	22.8999	22.8620	6.7144	6.2737	6.2747	26.4755	26.5794	16.0429	10.16
BEVSpread		53.9274	46.4839	46.2409	11.5392	10.8755	10.8755	27.4744	27.1533	23.0330	10.17
HeightFormer		27.6363	20.0780	20.0047	9.2197	8.8998	8.1294	14.8091	14.7369	8.8614	8.67
BEVHeight++		55.8025	47.1745	47.1523	8.8375	8.5998	8.7898	32.1752	31.8063	28.1942	7.79
Our		**72.6160**	**64.0757**	**64.1666**	**13.1446**	**12.5199**	**12.6940**	**35.9725**	**35.3335**	**31.5858**	17.66

Table 2: Comparison on the Rope3D [30] validation set under extrinsic perturbations.

Method	Deviation	$\text{Car}_{IoU=0.5}$ ↑			$\text{Big Vehicle}_{IoU=0.5}$ ↑		
		easy	moderate	hard	easy	moderate	hard
BEVHeight	yaw	43.9457	39.1533	39.0536	47.5926	46.5524	46.5961
BEVHeight++		50.3703	47.4298	45.1325	47.2277	44.3472	44.3624
BEVSpread		41.4254	35.4885	33.2370	41.5304	34.8029	34.6859
Our		**65.3942**	**62.3413**	**60.5443**	**54.5881**	**56.2474**	**56.2737**
BEVHeight	z	46.3001	44.7140	44.7303	40.4384	41.2793	41.3240
BEVHeight++		50.7617	35.2466	32.9790	42.0146	35.1928	35.0993
BEVSpread		41.8654	32.0237	32.0219	32.3425	32.2786	32.3853
Our		**65.6319**	**60.4119**	**58.2793**	**51.5508**	**52.6199**	**52.6557**

Perturbations details. We report robustness results under yaw rotation and z-axis translation, since these deviations are common in roadside deployments and induce direct range-dependent BEV misalignment [21,34]. Although RECO predicts full 6-DoF corrections, additional evaluations under pitch/roll perturbations are provided in the supplementary material for completeness.

Results under transient extrinsic perturbations. We compare RECO with other BEV-based methods like BEVHeight [29] and calibration-free methods like CBR [2] on the DAIR-V2X-I val set. As can be seen from Table 1, RECO achieves the best AP for car across all splits, improving over the strongest baseline by 0.77%, 1.22%, and 1.15%, and also yields consistent gains for pedestrian and cyclist. Under z perturbations, RECO surpasses the best baseline by a significant margin of 14.04%, 11.20%, 11.08% on Car, and improves pedestrian by about 1.61% to 1.82%. For cyclists, RECO is competitive and is notably stronger on the hard split, while CoBEV is slightly higher on easy/moderate.

Table 3: Robustness to persistent extrinsic deviations. All methods are trained with transient noise sampled from $\mathcal{N}(0, 0.5)$ and evaluated AP of car ($IoU = 0.5$, easy) under persistent deviations applied to either yaw rotation or z-axis translation. The columns labeled $(\mu, 0.5)$ denote evaluation with a mean shift $\mu \in \{-2, -1, 0, 1, 2\}$, while keeping the same noise scale (0.5). "↑" indicates higher is better.

Method	Deviation	$(-2, 0.5)\uparrow$	$(-1, 0.5)\uparrow$	$(0, 0.5)\uparrow$	$(1, 0.5)\uparrow$	$(2, 0.5)\uparrow$
BEVHeight	yaw	18.0282	56.7656	67.5595	51.4496	10.2972
CoBEV		10.0174	21.8748	31.4301	27.4911	6.434
HeightFormer		3.3049	12.7204	38.6363	10.9978	5.4677
BEVHeight++		13.7437	54.039	69.6915	52.6252	16.4732
Our		**49.3382**	**70.0774**	**70.4639**	**70.1092**	**44.3605**
BEVHeight	z	5.307	16.7973	48.3313	16.196	4.6595
CoBEV		12.7873	24.6159	46.0432	24.9035	13.2877
HeightFormer		3.2845	15.1936	27.6363	13.7415	2.9499
BEVHeight++		26.8912	**42.5294**	55.8025	45.6132	31.2033
Our		**31.0973**	37.3848	**72.616**	**64.3633**	**57.2835**

On the Rope3D dataset, we reproduce the BEV-based methods listed in Table 2 and compare our RECO against them. Under yaw jitter, RECO exceeds the best BEV baseline by about 15% across splits, and boosts big vehicle to 54.59%, 56.25%, and 56.27%. Under z perturbations, RECO again dominates, reaching 65.63%, 60.41%, 58.28% on car, with improvements of roughly 13.55% to 15.70% AP over the strongest baseline depending on the split.

Results of persistent extrinsic perturbations. Table 3 evaluates cross-condition generalization by training with transient extrinsic jitter and testing under persistent mean-shifted deviations. All models are trained with zero-mean Gaussian noise $\mathcal{N}(0, 0.5)$ to mimic short-term disturbances and are then evaluated on car AP ($IoU = 0.5$, easy) with a mean shift $\mu \in \{-2, -1, 0, 1, 2\}$ applied to yaw rotation or z-axis translation. Under this setting, our method achieves the best performance across both yaw and z perturbations, demonstrating improved robustness when models trained for transient noise are deployed under persistent miscalibration.

We observe a clear asymmetry between the yaw and z deviations. For yaw rotation, the performance under positive and negative mean shifts is approximately symmetric, suggesting that the detector's robustness is largely invariant to left–right perturbations. In contrast, for translation along the z-axis, the model exhibits higher tolerance to positive shifts than to negative ones. This behavior is physically plausible: z-translation changes the effective camera height and thus alters the apparent object scale and depth distribution in a direction-dependent manner, whereas yaw rotation induces a more symmetric left-right reprojection effect.

Results on rotational robustness. To evaluate robustness against different perturbations, we test all methods trained under yaw perturbation $\mathcal{N}(0, 0.5)$ on

Table 4: Comparison of different rotations on the DAIR-V2X-I.

Method	Deviation	Car$_{IoU=0.5}$ ↑			Ped.$_{IoU=0.25}$ ↑			Cyc.$_{IoU=0.25}$ ↑		
		easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
BEVHeight	pitch	53.1881	46.2226	46.3931	**9.3655**	**9.1158**	**9.2308**	29.3548	**29.8336**	**25.2666**
BEVHeight++	pitch	50.5844	42.2217	42.1829	7.6196	7.3630	7.3657	27.0724	26.7482	23.2941
Our	pitch	**65.3235**	**57.1345**	**57.9029**	8.4951	8.1553	8.2599	**30.2731**	29.6824	25.1420
BEVHeight	roll	52.3127	46.8362	47.0078	6.0710	5.6716	5.708	27.0194	26.4693	19.7767
BEVHeight++	roll	50.4606	42.1012	42.0723	6.7628	6.6048	6.2304	29.1563	**29.8950**	**23.0539**
Our	roll	**62.3741**	**54.0858**	**54.1871**	**6.9194**	**6.6530**	**6.4159**	**29.2147**	27.7538	22.7853
BEVHeight	pitch & roll	51.0522	44.2283	44.3883	5.7002	5.3966	5.4665	25.4191	24.9127	19.5592
BEVHeight++	pitch & roll	49.7541	41.4667	41.4342	**5.8089**	**5.6719**	**5.6707**	27.0194	26.4693	19.7767
Our	pitch & roll	**58.8318**	**50.8446**	**50.9040**	5.7964	5.5653	5.6622	**27.4047**	**26.7902**	**20.8360**

Table 5: Ablation study of compensation strategies under persistent extrinsic deviations. All variants are trained under transient noise $\mathcal{N}(0, 0.5)$ and evaluated with mean-shifted deviations $(\mu, 0.5)$, where $\mu \in \{-2, -1, 0, 1, 2\}$, applied to either yaw rotation or z-axis translation. We compare No_Comp, Global_Comp, Hard_Comp (range-aware compensation with hard assignment), and Soft_Comp (range-aware compensation with soft gating). "↑" indicates higher is better.

Method	Deviation	$(-2, 0.5)$ ↑	$(-1, 0.5)$ ↑	$(0, 0.5)$ ↑	$(1, 0.5)$ ↑	$(2, 0.5)$ ↑
No_Comp	yaw	18.0282	56.7656	67.5595	51.4496	10.2972
Global_Comp	yaw	32.2572	33.2531	33.2416	33.2421	33.0389
Hard_Comp	yaw	40.0773	64.9268	68.0451	65.7611	40.0988
Soft_Comp	yaw	**49.3382**	**70.0774**	**70.4639**	**70.1092**	**44.3605**
No_Comp	z	5.307	16.7973	48.3313	16.196	4.6595
Global_Comp	z	23.0018	23.2671	23.2683	23.3090	23.2668
Hard_Comp	z	**31.6210**	36.8859	68.3201	**65.0907**	56.7534
Soft_Comp	z	31.0973	**37.3848**	**72.616**	64.3633	**57.2835**

pitch, roll, and joint pitch&roll perturbations, without retraining or fine-tuning. Table 4 reports the comparison results on DAIR-V2X-I. Following BEVHeight++ [28], the angles sampled from $\mathcal{N}(1, 1.67)$ are applied to the camera pitch and roll at test time. Our method outperforms comparative methods, with gains over BEVHeight [29] under pitch perturbations reaching +12.14%, +10.91%, +11.51% in the car category. For the cyclist, our method also provides competitive performance, achieving 30.27%, 29.68%, 25.14% AP under pitch deviation and 27.40%, 26.79%, 20.84% under joint pitch&roll deviation, both better than BEVHeight++. Overall, these results show that our method provides markedly better robustness to rotational extrinsic perturbations, especially for car and cyclist detection.

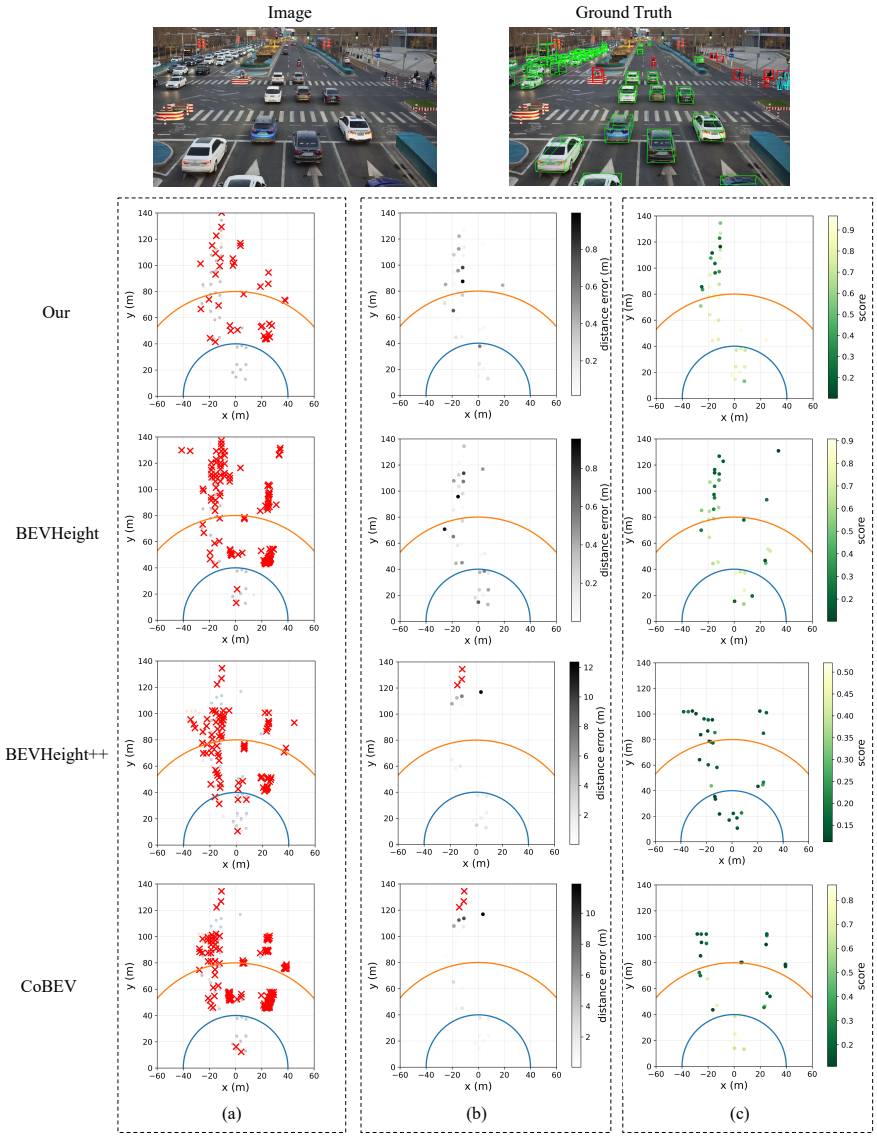

Fig. 5: Qualitative comparison under extrinsic perturbations. **Top**: a representative roadside image (left) and the corresponding ground-truth annotations (right). **Bottom**: BEV visualizations for different methods (rows). (a) Missed and false detections: red crosses denote missed ground-truth objects and unmatched predictions, while dots indicate ground-truth and matched predictions. (b) Distance error map for matched predictions: darker colors denote larger distance errors between predicted and ground-truth, while red crosses indicate missed ground-truth. (c) Confidence map of predictions: brighter colors indicate higher detection confidence. The blue and orange semi-circles indicate range rings.

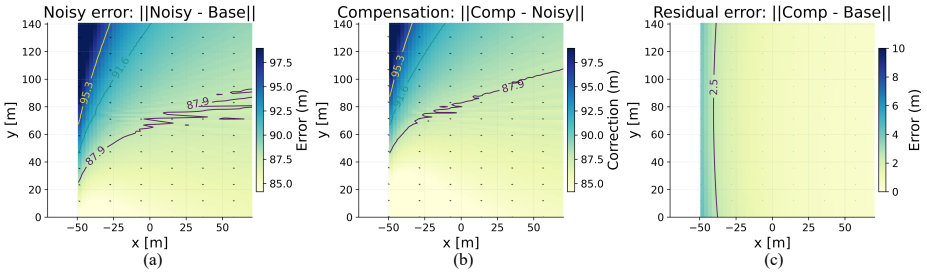

Fig. 6: Visualization of BEV grid displacement and compensation under yaw $\mathcal{N}(0, 0.5)$. (a) Displacement magnitude after injecting extrinsic noise, measured as $||\mathbf{G}_{\text{noisy}} - \mathbf{G}_{\text{base}}||$ at each BEV cell. (b) The compensation applied by our method, $||\mathbf{G}_{\text{comp}} - \mathbf{G}_{\text{noisy}}||$, shows how the geometry field is corrected across the grid. (c) Residual displacement after compensation, $||\mathbf{G}_{\text{comp}} - \mathbf{G}_{\text{base}}||$. The overlaid polylines are iso-contours of the plotted quantity. Colorbars indicate the displacement/correction magnitude in meters.

4.3 Ablation Study

Table 5 demonstrates the effectiveness of introducing compensation into the projection pipeline under persistent mean-shifted extrinsic deviations. Compared with No_Comp, adding a Global_Comp mechanism substantially improves robustness in several shifted settings, indicating that a single global 6-DoF correction can partially alleviate projection misalignment under persistent drift. More importantly, range-aware compensation yields consistent additional gains over global compensation. Both Hard_Comp and Soft_Comp significantly improve performance under large mean shifts, and Soft_Comp performs best overall across yaw and z perturbations. This suggests that range-adaptive modeling is beneficial and that differentiable soft gating mitigates the discontinuities of hard partitioning, leading to stronger robustness.

4.4 Visualization Results

Detection result. Fig. 5 provides a qualitative comparison of RECO and prior methods under the same scene and yaw deviations $\mathcal{N}(0, 0.5)$. As shown in Fig. 5(a), prior methods exhibit noticeably more false positives or missing ground-truth boxes, especially in the far-range region, while RECO yields fewer failure cases. The error map in Fig. 5(b) further shows that RECO reduces distance errors for matched detections, with fewer high-error points in the far-range region, suggesting improved detection effectiveness under extrinsic perturbations. In addition, Fig. 5(c) shows that RECO produces more reliable confidence estimates, assigning higher scores to correct detections.

Compensation result. Fig. 6 provides direct evidence that the proposed compensation restores the consistency of BEV sampling under extrinsic perturbations. The noise error displacement concentrates and grows with range (Fig. 6(a)),

indicating a highly non-uniform distortion of the geometry field rather than a simple global shift. Our method produces a correction that closely counteracts this distortion, yielding near-zero residual displacement over most of the BEV plane (Fig. 6(b) and Fig. 6(c)). This shows that RECO effectively cancels the geometric misalignment introduced by pose noise and recovers an approximately noise-free projection geometry.

5 Conclusion

In this paper, we address the long-standing sensitivity of roadside monocular 3D detection to camera extrinsic drift in real deployments. We presented RECO, a region-aware extrinsic compensation framework that predicts a learnable range boundary and estimates near/far 6-DoF pose corrections to refine nominal extrinsics. To avoid non-smooth BEV sampling induced by region-aware offsets, we further introduced a differentiable soft gated compensation projection that smoothly blends the compensated geometries and enables stable optimization with reprojection supervision. Extensive experiments on DAIR-V2X-I and Rope3D demonstrate that RECO consistently improves robustness under both transient jitter and persistent deviations, while maintaining competitive detection accuracy. Future work will explore extending this region-aware compensation mechanism to multi-camera infrastructure networks and V2I cooperative perception systems to ensure broader global consistency.

We hope this work encourages future research on reliable infrastructure perception under imperfect calibration and facilitates the practical deployment of roadside vision systems in the wild.

6 Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No. 2023YFB4301900), by the Shenzhen Key Industry Research Program (ZDCY20250901112501002), by the Shenzhen Natural Science Foundation General Research Project (JCYJ20240813151445059), and by the Guangdong Basic and Applied Basic Research Foundation (grant No.2026A1515010917).

References

1. Aoki, Y., Goforth, H., Srivatsan, R.A., Lucey, S.: Pointnetlk: Robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7163–7172 (2019)
2. Fan, S., Wang, Z., Huo, X., Wang, Y., Liu, J.: Calibration-free bev representation for infrastructure perception. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9008–9013 (2023). https://doi.org/10.1109/IROS55552.2023.10341916

3. Fu, K., Liu, S., Luo, X., Wang, M.: Robust point cloud registration framework based on deep graph matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8893–8902 (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arxiv 2022. arXiv preprint arXiv:2203.17054
6. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
7. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4267–4276 (2021)
8. Iyer, G., Ram, R.K., Murthy, J.K., Krishna, K.M.: Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1110–1117. IEEE (2018)
9. Ji, Y., Zhou, Z., Yang, Z., Huang, Y., Zhang, Y., Zhang, W., Xiong, L., Yu, Z.: Toward autonomous vehicles: A survey on cooperative vehicle-infrastructure system. Iscience **27**(5) (2024)
10. Li, J., Lee, G.H.: Deepi2p: Image-to-point cloud registration via deep classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15960–15969 (2021)
11. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 1477–1485 (2023)
12. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2020–2036 (2024)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Lv, X., Wang, B., Dou, Z., Ye, D., Wang, S.: Lccnet: Lidar and camera self-calibration using cost volume network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2894–2901 (2021)
15. Ma, Y., Li, Y., Hua, W., Kong, S.: Roadside monocular 3d detection prompted by 2d detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2026)
16. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European conference on computer vision. pp. 194–210. Springer (2020)
17. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11143–11152 (2022)
18. Ren, S., Zeng, Y., Hou, J., Chen, X.: Corri2p: Deep image-to-point cloud registration via dense correspondence. IEEE Transactions on Circuits and Systems for Video Technology **33**(3), 1198–1208 (2022)
19. Schneider, N., Piewak, F., Stiller, C., Franke, U.: Regnet: Multimodal sensor registration using deep neural networks. In: 2017 IEEE intelligent vehicles symposium (IV). pp. 1803–1810. IEEE (2017)

20. Scott, W.R., Roth, G., Rivest, J.F.: Pose error effects on range sensing. National Research Council of Canada (2002)
21. Shi, H., Pang, C., Zhang, J., Yang, K., Wu, Y., Ni, H., Lin, Y., Stiefelhagen, R., Wang, K.: Cobev: Elevating roadside 3d object detection with depth and height complementarity. IEEE Transactions on Image Processing (2024)
22. Vuong, K., Tamburo, R., Narasimhan, S.G.: Toward planet-wide traffic camera calibration. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8553–8562 (2024)
23. Wang, W., Lu, Y., Zheng, G., Zhan, S., Ye, X., Tan, Z., Wang, J., Wang, G., Li, X.: Bevspread: Spread voxel pooling for bird's-eye-view representation in vision-based roadside 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14718–14727 (2024)
24. Wang, Y., Solomon, J.M.: Deep closest point: Learning representations for point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3523–3532 (2019)
25. Wu, J., Yuan, M., Wang, T., Jia, X., Yan, D.M.: Heightformer: Single-imagery height estimation transformer with bilateral feature pyramid fusion. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)
26. Xia, Z.X., Fadadu, S., Shi, Y., Foucard, L.: Robust long-range perception against sensor misalignment in autonomous vehicles. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5761–5770. IEEE (2025)
27. Xiao, Y., Li, Y., Meng, C., Li, X., Ji, J., Zhang, Y.: Calibformer: A transformer-based automatic lidar-camera calibration network. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 16714–16720. IEEE (2024)
28. Yang, L., Tang, T., Li, J., Yuan, K., Wu, K., Chen, P., Wang, L., Huang, Y., Li, L., Zhang, X., et al.: Bevheight++: Toward robust visual centric 3d object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
29. Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., Zhang, X., Chen, P.: Bevheight: A robust framework for vision-based roadside 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21611–21620 (2023)
30. Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E.: Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21341–21350 (2022)
31. Yew, Z.J., Lee, G.H.: Rpm-net: Robust point matching using learned features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11824–11833 (2020)
32. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., et al.: Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21361–21370 (2022)
33. Zhang, W., Gao, Y., Jiang, Z., Mao, R., Zhou, S.: Calibration-free roadside bev perception with v2x-enabled vehicle position assistance. Sensors **25**(13) (2025), https://www.mdpi.com/1424-8220/25/13/3919
34. Zhu, Y., Wang, Z., Wang, Y.: Mamv2xcalib: V2x-based target-less infrastructure camera calibration with state space model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26696–26705 (2025)