# Layout-Robust LiDAR 3D Object Detection via Multi-Representation Fusion

Wenkai Zhu[1], Xu Li[1], Wang Xu[2], Linru Li[3], Longjie Liao[1], Benwu Wang[1], Xueliang Ren[1], Xiaoyu Yue[1], JiXian Zheng[1], Jinfeng Wu[1]

[1]Southeast University, [2]Shanghai Jiao Tong University, [3]Nanjing Normal University
[4] Momenta.ai

## Abstract

*3D object detection is essential for autonomous driving, yet LiDAR-based detectors often generalize poorly across different vehicle platforms due to heterogeneous multi-LiDAR system architectures. Variations in sensor-suite configurations—such as differences in the number, placement, and orientation of LiDAR units—produce distinct, non-transferable environmental representations. As a result, model performance can degrade significantly when encountering unseen system layouts. To address this challenge, we propose a unified representation framework that remains robust under diverse LiDAR placements. The architecture consists of two core components: (1) a Multi-View Fusion (MVF) module that aggregates complementary information through point–voxel attention to learn a unified, view-invariant representation and mitigate feature imbalance under uneven point density, and (2) a Motion-Guided Spatial–Temporal Fusion (MG-STF) module that leverages motion cues to align spatial–temporal features across consecutive frames. Experiments on the public Track 3 benchmark demonstrate competitive performance, with our approach ranking 5th on the leaderboard. Ablation studies further validate the effectiveness of both modules in improving generalization under density imbalance and layout variation.*

***Keywords:*** *3D object detection, LiDAR, multi-view fusion, motion-guided spatial–temporal fusion.*

## 1. Introduction

LiDAR-based 3D object detection has become central to autonomous-driving perception[1–14]. By providing metrically accurate range under diverse lighting and weather, LiDAR supports core tasks—including odometry and mapping [15–17], multi-object tracking [18–20], and detection [20–24]. Yet LiDAR returns are inherently unordered, sparse, and non-uniform [25–28], which complicates feature extraction and prevents the direct transfer of image-centric pipelines. These characteristics motivate representations and learning frameworks tailored specifically to point-cloud data.

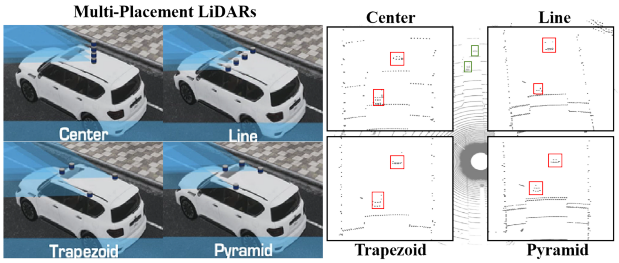


Figure 1. The right subfigure visualizes point-density imbalance for the same frame under different LiDAR configurations. As observed, the center layout yields more uniform sampling, whereas alternative layouts exhibit sector-level voids, anisotropic sparsity, and vertical stratification—particularly along the vehicle flanks and at long range—accompanied by near-field over-sampling, far-field thinning, and intensified self-occlusion. These effects underscore pronounced density nonuniformity and view-induced bias.

However, different combinations of LiDAR mounting positions cause the cross-platform generalization of current 3D object detectors for autonomous driving to remain limited [25, 29–34]. Heterogeneity in mounting height, pitch/yaw, and single- versus multi-sensor configurations alters vantage point, field-of-view overlap, and occlusion patterns, thereby inducing systematic distribution shifts—such as range-dependent density decay, vertical stratification, and anisotropic sparsity [26, 35–42].

Models often tacitly overfit the training set's sampling geometry and hyperparameters (*e.g.*, voxel granularity and receptive-field design), yielding layout-sensitive rather than scene-invariant representations and decision boundaries [43–45]. As illustrated in Fig. 1, center-mounted and linear layouts around the ego vehicle induce pronounced density disparities; consequently, detectors relying on a single representation tend to misalign when transferred to unseen layouts, exhibiting weakened generalization and substantial performance degradation.

Thus, single-representation 3D detectors generalize un-

reliably in the face of heterogeneous LiDAR mounting schemes and density drift. Current LiDAR-based 3D detectors largely fall into two families: voxel-based and point-based. Voxel methods discretize irregular point clouds into grids and learn features with 3D convolutions [20, 43, 46–51]; accuracy improves with finer voxels but at substantial computational cost, whereas coarser voxels are efficient yet blur local structure. Point methods process raw points and extract keypoint features via PointNet and its variants [52, 53] and representative point-based detectors [54–56]; although they avoid quantization loss, efficiently capturing fine-grained geometry under sparse, irregular sampling remains challenging.

To address the aforementioned cross-layout generalization limitations caused by heterogeneous LiDAR placements, we propose a multi-view unified-representation 3D object detection framework comprising two key modules: Multi-Representation Fusion (MRF) and Motion-Guided Spatio-Temporal Fusion (MG-STF). MRF performs density-adaptive fusion across Point, Voxel, and BEV views via point–voxel attention, while MG-STF leverages motion cues to align spatio-temporal features across adjacent frames.

Motivated by insights from Place3D [57] on layout-sensitive perception, we develop a cross-layout–robust approach and implement it in MMDetection3D [58] with PV-RCNN+ as the baseline [11, 12]. For fair comparison, we keep the backbone, data preprocessing, and training schedule unchanged, and introduce two modules only: MVF and MG-STF. MVF performs density-adaptive fusion across Point/Voxel/BEV views via point–voxel attention [59], while MG-STF leverages motion priors to achieve spatiotemporal feature alignment across frames [18, 60]. These incremental additions preserve apples-to-apples fairness with the baseline and deliver notable gains in generalization under heterogeneous LiDAR layouts.

## 2. Related Work

Research on LiDAR-based 3D object detection primarily follows two paradigms: single-representation approaches and point–voxel representation–based approaches.

### 2.1. Single Representation-Based 3D Detection

Single-representation LiDAR detectors are largely split into voxelized and point-centric paradigms. Voxelized methods discretize irregular clouds into grids to enable 3D CNNs (*e.g.*, VoxelNet) [47]; computation is alleviated by sparse convolutions [26, 48] and by BEV/Pillar abstractions that reuse 2D CNNs for real-time inference [20, 43, 61]. In autonomous driving, these systems are widely benchmarked on nuScenes [1]. Point-centric methods operate directly on raw points with permutation-invariant encoders (PointNet/PointNet++) [52, 53]; two-stage or refined pipelines such as PointRCNN, PointGNN, and 3DSSD improve localization via proposals, graph reasoning, or single-stage keypoint sampling [54–56, 62, 63]. These approaches better preserve fine geometry and often achieve competitive or superior accuracy, but efficient local aggregation under severe sparsity remains challenging.

### 2.2. Point-Voxel Representation-Based 3D Detection

A complementary line of work unifies point- and voxel-based paradigms to couple point-level fidelity with voxel/BEV efficiency, exemplified by PV-RCNN and its successor PV-RCNN++ [11, 12]. PV-RCNN generates proposals via sparse 3D convolutions and refines them by lifting multi-scale voxel cues to a sparse set of keypoints, while PV-RCNN++ enhances local vector representation and end-to-end optimization [12]. These hybrids are commonly evaluated for autonomous-driving detection on nuScenes [1]. Building on this blueprint, PVAFN preserves fine-grained point features end-to-end and fuses them with voxel–BEV context through a learned attention mechanism, enabling density-adaptive, bidirectional alignment between representations; this tightens point–voxel integration and mitigates quantization artifacts while retaining the computational advantages of voxelized processing.

## 3. Methodology

### 3.1. System Overview

To mitigate the cross-layout distribution shift caused by heterogeneous LiDAR placements, we propose a unified 3D detection framework with two plug-and-play modules: Multi-Representations Fusion Module (MRFM) and Motion-Guided Spatial–Temporal Fusion (MG-STF).

As shown in (Fig. 2), raw point clouds are encoded into three complementary views—Point, Voxel, and BEV. MRFM performs Representations-wise self-attention followed by point–voxel cross-attention to adaptively fuse the three views into a per-frame representation $\mathbf{f}_{pbv}^{t}$ that is less sensitive to sampling geometry. MG-STF then exploits motion cues to align adjacent frames: it forms a temporal residual between $(\mathbf{f}_{pbv}^{t}, \mathbf{f}_{pbv}^{t-1})$, refines it with light-weight attention to obtain motion guidance, and injects this guidance into a fusion Transformer to produce the final predictions.

### 3.2. Multi-Representations Fusion Module

To address the inconsistencies and misalignment across 3D representations—where points are sparse while voxels and BEV features are dense, with differing receptive fields and noise statistics. we leverage a Multi-Representation Fusion Module (MRFM). The MRFM takes point features $\mathbf{f}_p$, voxel features $\mathbf{f}_v$, and BEV features $\mathbf{f}_B$ as inputs. It first applies self-attention to each representation separately to capture local–global context:

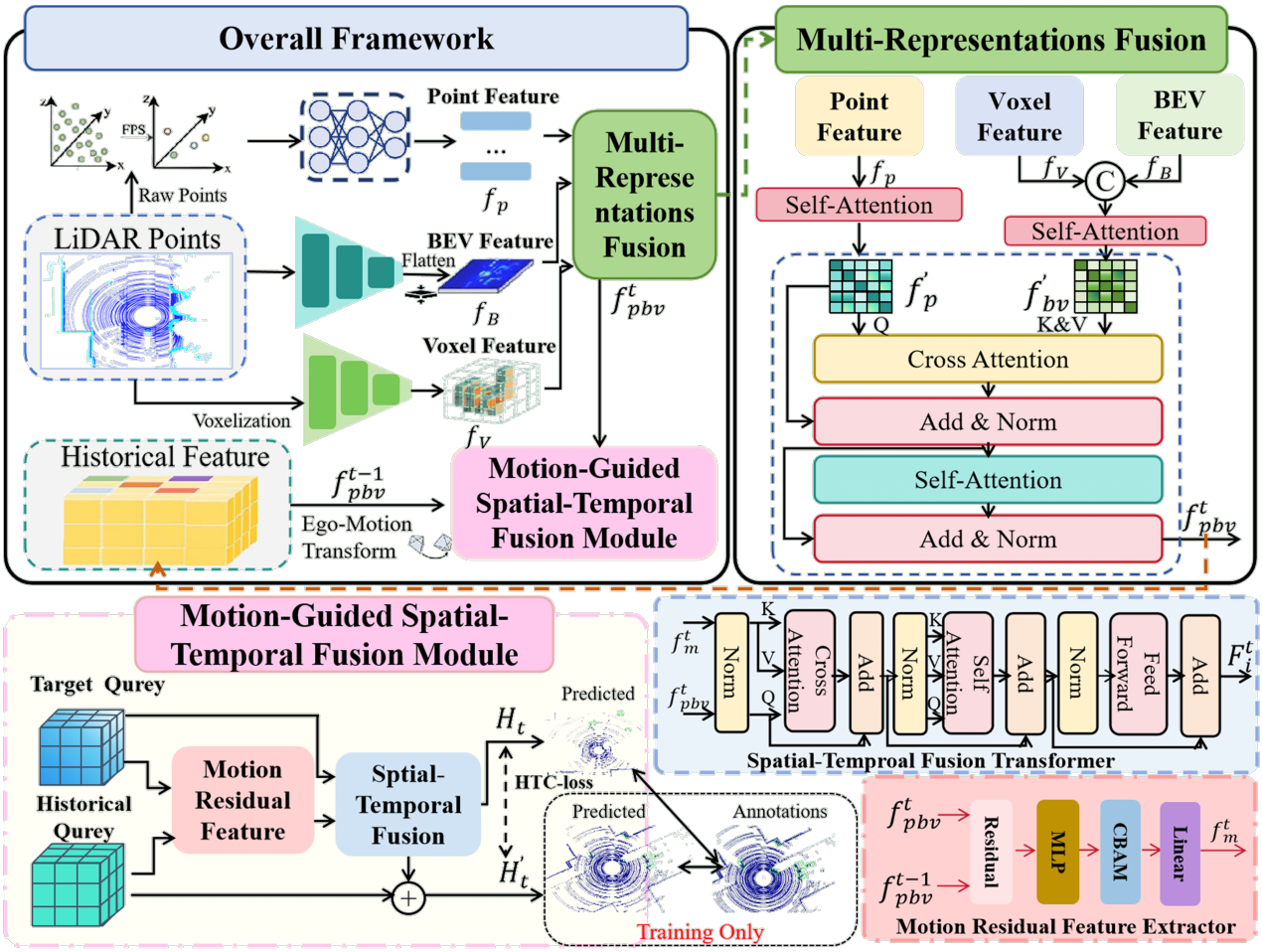


Figure 2. Overview of the framework. The **Multi-representation Fusion Module** (MRFM) fuses Point ($\mathbf{f}_p$), Voxel ($\mathbf{f}_v$), and BEV ($\mathbf{f}_B$) features via view-wise self-attention and point–voxel cross-attention to produce the per-frame representation $\mathbf{f}^t_{pbv}$. The **Motion-Guided Spatial–Temporal Fusion** forms a temporal residual from $(\mathbf{f}^t_{pbv}, \mathbf{f}^{t-1}_{pbv})$ and injects it into a fusion Transformer to yield the final prediction $\mathbf{H}_t$.

$$\tilde{\mathbf{f}}p = \mathrm{SA}(\mathbf{f}p), \qquad \tilde{\mathbf{f}}bv = \mathrm{SA}\big([\mathbf{f}v|\mathbf{f}_B]\big), \tag{1}$$

where $\mathrm{SA}(\cdot)$ denotes self-attention and $[\cdot|\cdot]$ denotes concatenation. The concatenated voxel–BEV tensor $\tilde{\mathbf{f}}_{bv}$ serves as keys and values in a Point–Voxel Cross-Attention module, with point features $\tilde{\mathbf{f}}_p$ as queries, to align the heterogeneous representations:

$$\mathbf{Q} = \mathbf{W}q\tilde{\mathbf{f}}p, \quad \mathbf{K} = \mathbf{W}k\tilde{\mathbf{f}}bv, \quad \mathbf{V} = \mathbf{W}v\tilde{\mathbf{f}}bv, \tag{2}$$

$$\hat{\mathbf{f}}_p = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{!\top}}{\sqrt{d}}\right)\mathbf{V}. \tag{3}$$

The output is further refined via a Transformer block with residual connections and layer normalization (LN), yielding the final fused representation:

$$\mathbf{f}^tpbv = \mathrm{LN}\Big(\mathrm{SA}\big(\mathrm{LN}(\tilde{\mathbf{f}}p + \hat{\mathbf{f}}p)\big) + \mathrm{LN}(\tilde{\mathbf{f}}p + \hat{\mathbf{f}}_p)\Big), \tag{4}$$

where $d$ is the channel dimension, $\hat{\mathbf{f}}p$ is the cross-representation aligned point feature, and $\mathbf{f}^tpbv$ is the resulting robust fused representation. Thanks to this cross-representation alignment, $\mathbf{f}^t_{pbv}$ exhibits strong robustness to heterogeneous multi-LiDAR configurations.

### 3.3. Motion-Guided Spatio-Temporal Fusion Module

Given the fused per-frame representation $\mathbf{f}^t_{pbv}$ (from the multi-representation backbone) and its previous-step counterpart $\mathbf{f}^{t-1}_{pbv}$, we introduce a *Motion-Guided Spatial–Temporal Fusion Module* composed of two parts: (i) a Motion Residual Feature Extractor; (ii) a Spatial–Temporal Fusion Trans-

Table 1. Per-class results (AP / ATE / ASE / AOE / AVE / AAE).

| | AP↑ | | ATE↓ | | ASE↓ | | AOE↓ | | AVE↓ | | AAE↓ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | Ours | BEV-L | Ours | BEV-L | Ours | BEV-L | Ours | BEV-L | Ours | BEV-L | Ours | BEV-L |
| car | 0.726 | 0.698 | 0.164 | 0.159 | 0.162 | 0.136 | 1.119 | 1.074 | 0.885 | 1.954 | 0.064 | 0.241 |
| truck | 0.756 | 0.742 | 0.143 | 0.119 | 0.126 | 0.069 | 1.152 | 1.042 | 1.153 | 2.505 | 0.067 | 0.296 |
| bus | 0.512 | 0.502 | 0.096 | 0.079 | 0.139 | 0.046 | 1.143 | 1.126 | 1.049 | 2.060 | 0.077 | 0.177 |
| pedestrian | 0.803 | 0.560 | 0.082 | 0.118 | 0.139 | 0.112 | 0.940 | 1.160 | 0.314 | 0.851 | 0.042 | 0.353 |
| motorcycle | 0.791 | 0.737 | 0.100 | 0.087 | 0.135 | 0.091 | 1.119 | 1.323 | 0.610 | 2.161 | 0.065 | 0.414 |
| bicycle | 0.756 | 0.655 | 0.072 | 0.067 | 0.179 | 0.133 | 1.181 | 1.078 | 0.476 | 1.188 | 0.056 | 0.248 |

former. The overall goal is to inject motion cues into temporal fusion while preserving spatial details.

#### 3.3.1 Motion Residual Feature Extractor.

The Motion Residual Feature Extractor derives explicit motion guidance by computing the differential encoding between consecutive fused features $\mathbf{f}^t pbv$ and $\mathbf{f}^{t-1} pbv$:

$$\mathbf{r}^t = \delta\left(\mathbf{f}^t pbv, , \mathbf{f}^{t-1} pbv\right) = \mathbf{f}^t pbv - \mathbf{f}^{t-1} pbv. \quad (5)$$

The resulting temporal residual map $\mathbf{r}^t$ accentuates time-varying regions: static areas yield values near zero, while moving objects produce significant activations. This residual is then transformed by an MLP, refined by a lightweight CBAM gate to emphasize informative features, and linearly projected to produce the final motion-guidance feature $\mathbf{f}^t m$ for subsequent fusion:

$$\mathbf{f}^t m = \mathbf{W}_m \cdot \mathrm{CA}\big(\mathrm{MLP}(\mathbf{r}^t)\big). \quad (6)$$

#### 3.3.2 Spatial-Temporal Fusion Transformer.

We fuse the current target query $\mathbf{f}^t pbv$ with the historical query $\mathbf{f}^{t-1} pbv$ under the motion guidance of $\mathbf{f}_m^t$. The process, termed *Motion-Guided Fusion*, can be compactly expressed as:

$$\mathbf{F}^t = \mathrm{TF}\left(\mathrm{CrossAttn}\left(\mathbf{f}^t pbv,\ [\mathbf{f}_m^t \| \mathbf{f}_m^t]\right)\right) \quad (7)$$

where the $\mathrm{CrossAttn}(\mathbf{Q}, \mathbf{KV})$ function denotes a multi-head cross-attention layer with pre-normalization and residual connections, and $\mathrm{TF}(\cdot)$ represents a standard encoder stack comprising self-attention and feed-forward layers.

where MHA, LN, and FFN denote multi-head attention, layer normalization, and a feed-forward network, respectively.

### 3.4. Overall Training Loss Function

Inspired by OnlineBEV [60], we leverage the HTC-loss to explicitly supervise the alignment between the current feature and the temporally aligned historical feature. This direct consistency constraint enhances the stability and coherence of the features, leading to more effective long-term temporal fusion.

The HTC-loss is computed by processing both $H_t$ and $\hat{H}_t$ through two shared-weight prediction heads to produce heatmaps $P_t$ and $\hat{P}t$, respectively. The loss is then defined as the L2 distance between them:

$$\mathcal{L}\mathrm{cons} = \|P_t - \hat{P}_t\|_2^2. \quad (6)$$

A stop-gradient operation is applied to $P_t$ to ensure that the gradient only drives the alignment of $\hat{H}_t$ towards the target $H_t$.

The overall training objective is a weighted sum of task losses and our consistency loss:

$$\mathcal{L} = w_{\mathrm{cls}}\mathcal{L}\mathrm{cls} + w\mathrm{reg}\mathcal{L}\mathrm{reg} + w\mathrm{cons}\mathcal{L}\mathrm{cons}. \quad (8)$$

Here, the weights $w\mathrm{cls}$, $w_{\mathrm{reg}}$, and $w_{\mathrm{cons}}$ are set to 0.75, 0.25, and 1.5, respectively. The loss terms $\mathcal{L}\mathrm{cls}$, $\mathcal{L}\mathrm{reg}$, and $\mathcal{L}_{\mathrm{cons}}$ correspond to focal loss, L1 regression loss, and the HTC-loss, respectively.

## 4. Experiments

### 4.1. Dataset

We use the official data provided by the *RoboSense Challenge 2025* [65] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [66, 67] at ICRA 2023 and the *RoboDrive Challenge 2024* [68, 69] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in this competition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [57, 70–73]. Specifically, this task is built upon the **Place3D** dataset [57] in **Track 3**, which provides a standardized foundation

Table 2. Overall results on the benchmark. ↑: higher is better; ↓: lower is better.

| Method | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ | NDS↑ |
|---|---|---|---|---|---|---|---|
| BEVFusion-L [64] | 0.605 | 0.112 | 0.0102 | 1.134 | 1.786 | 0.288 | 0.575 |
| PointPillars [43] | 0.543 | 0.119 | 0.111 | 1.284 | 2.023 | 0.326 | 0.508 |
| CenterPoint [20] | 0.659 | 0.099 | 0.092 | 1.068 | 1.682 | 0.271 | 0.610 |
| Ours | 0.724 | 0.110 | 0.147 | 1.109 | 0.748 | 0.062 | 0.655 |

Table 3. Ablation on Place3D (Stage-1 val). MRFM focuses on Multi-representation fusion; MG-STF injects motion cues for spatial–temporal alignment. ↑: higher is better; ↓: lower is better.

| Variant | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ | NDS↑ |
|---|---|---|---|---|---|---|---|
| Backbone only | 0.597 | 0.102 | 0.120 | 1.140 | 1.520 | 0.210 | 0.620 |
| + MRFM | 0.679 (+0.082) | 0.104 | 0.128 | 1.115 | 1.450 | 0.190 | 0.635 |
| MRFM + MG-STF (Ours) | 0.724 (+0.127) | 0.110 | 0.147 | 1.109 | 0.748 | 0.062 | 0.655 |

for benchmarking performance under challenging conditions such as cross-domain shifts, sensor variability, and multi-modal alignment. Place3D comprises 11 LiDAR layouts in total—seven baseline layouts inspired by production AV rigs and four layouts obtained via sensor placement optimization.

*Stage-1 (train/val)* contains 200 scenes (8,000 frames) with six synchronized cameras and LiDAR captured from four distinct mounting positions; the official split is 125 scenes (5,000 frames) for training and 75 scenes (3,000 frames) for validation. This stage supports either single-layout training with cross-layout validation or joint training over all layouts.

*Stage-2 (test)* provides six-camera imagery together with LiDAR from six mounting layouts; all LiDAR files are placed in a single directory to simplify evaluation (no path switching is required). For more details, please refer to the corrseponding GitHub repositories.

## 4.2. Evaluation Metrics

We follow the official nuScenes[1, 74]evaluation protocol, which includes one precision-oriented metric and five error metrics, together with the composite NDS score. The mAP metric adopts center-distance matching rather than IoU; AP is computed for distance thresholds across all classes and then averaged. The five error terms quantify complementary aspects of detection quality: mATE (m) measures center translation error; mASE captures scale discrepancy as IoU after aligning position and yaw; mAOE (rad) measures orientation error; mAVE (m/s) measures velocity-vector error; and mAAE measures attribute misclassification (*e.g.*, moving vs. static). Lower values are better for all five. The nuScenes Detection Score (NDS) aggregates mAP and the five errors via fixed weights and threshold-based normalization, yielding a single score that reflects both detection precision and the quality of localization, pose, motion, and attribute estimation.

To evaluate our approach, we compare against three representative 3D object detection baselines: BEVFusion, PointPillars, and CenterPoint [20, 43, 64]. All methods are trained and evaluated under the same framework and protocol, with matched input preprocessing, training schedules, and post-processing to ensure a fair comparison.

## 4.3. Implementation Details

We implement our model on MMDetection3D [75]. The point encoder follows PointNet++ set-abstraction with farthest-point sampling and ball-query grouping, aggregating multi-scale neighborhood features with shared MLPs to form keypoint descriptors [53]. The voxel encoder first voxelizes points (HardVFE/DynamicVFE) and computes per-voxel statistics via point-wise MLP + pooling, which are then processed by a sparse 3D CNN backbone (SECOND/Minkowski-style) to learn volumetric features [26, 48]. For BEV processing, we either collapse the height dimension of sparse 3D features or scatter pillar features to a 2D canvas to obtain a BEV "pseudo-image" followed by a 2D backbone/neck (*e.g.*, SECONDFPN) for multi-scale fusion and detection [43]. Detection heads (center-based or anchor-based) operate on the BEV map, while optional RoI modules pool point/voxel features around proposals for refinement.

## 4.4. Comparative Study

### 4.4.1 Overall comparison.

Table 2 reports the aggregate metrics: our model achieves the best *mAP* (0.724) and *NDS* (0.655), outperforming BEVFusion-L by +12.1% mAP and +13.9% NDS, and exceeding CenterPoint by +5.1% and +7.4%, respectively.

These gains are attributed to the proposed MRFM, which reconciles Point/Voxel/BEV discrepancies via representation-wise self-attention and point–voxel cross-attention to mitigate view-induced bias and density nonuniformity—thereby strengthening per-frame semantics and slightly improving orientation (AOE: $-2.2\%$ vs. BEVFusion-L)—and to MG-STF, which injects motion priors through the temporal residual pathway and fuses them with cross-attention in the STFT block, leading to pronounced reductions in velocity and attribute errors (mAVE $0.748$ vs. $1.786$, $-58.1\%$; mAAE $0.062$ vs. $0.288$, $-78.5\%$) that translate directly into higher NDS.

#### 4.4.2 Per-class analysis.

From Table 1, AP improves across all categories compared with BEVFusion-L, with the largest gains on dynamic classes—pedestrian (+0.243, +43.4%), motorcycle (+0.054, +7.3%), and bicycle (+0.101, +15.4%). These improvements are consistent with the role of MG-STF: the temporal residual $\mathbf{r}^t$ and motion descriptor $\mathbf{f}_m^t$ selectively emphasize time-varying tokens, leading to pronounced per-class reductions in AVE (*e.g.*, pedestrian $-63.1\%$, motorcycle $-71.8\%$, bicycle $-59.9\%$) and concomitant AOE decreases, which translate into disproportionate AP/NDS gains for dynamic objects. Meanwhile, MRFM aligns sparse point cues with voxel/BEV context, improving recall for thin and anisotropic structures (bicycle/motorcycle) despite layout-induced sparsity, thereby contributing to the across-the-board AP uplift.

### 4.5. Ablation Study

Starting from the backbone-only model (mAP/NDS $= 0.597/0.620$), introducing **MRFM** lifts mAP to $0.679$ (+13.7% rel.) and NDS to $0.635$ (+2.4%), while modestly improving orientation (AOE $1.140 \rightarrow 1.115$, $-2.2\%$) and slightly reducing motion/attribute errors (AVE $1.520 \rightarrow 1.450$, AAE $0.210 \rightarrow 0.190$). The increases in ATE/ASE (0.102→0.104, 0.120→0.128) indicate that cross-representation mixing can dilute voxel-dominant size cues even as per-frame semantics strengthen—consistent with MRFM's role in mitigating view-induced bias and density nonuniformity. Appending **MG-STF** on top of MRFM (Ours) further raises mAP/NDS to $0.724/0.655$ and drives large reductions in AVE/AAE (1.450→0.748, $-48.4\%$; 0.190→0.062, $-67.4\%$) by injecting temporal residuals and motion guidance into the fusion Transformer.

## 5. Conclusion

We addressed cross-layout generalization in LiDAR 3D detection with two plug-and-play modules: the Multi-Representation Fusion Module (MRFM) for point/voxel/BEV reconciliation and the Motion-Guided Spatio-Temporal Fusion (MG-STF) for motion-aware temporal alignment. On Place3D (ROBOSENSE protocol), our model attains 0.724 mAP and 0.655 NDS, surpassing BEVFusion-L by +0.075 mAP (+11.6%) and +0.080 NDS (+13.9%). Future work will decouple geometry-specific supervision and add scale-calibration priors to tighten box size and localization without sacrificing temporal robustness.

## References


[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venıce Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 1, 2, 5

[2] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuScenes: A large-scale benchmark for LiDAR panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[4] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020.

[5] Xiang Xu et al. Beyond one shot, beyond one perspective: Cross-view and long-horizon distillation for better LiDAR representations. In *IEEE/CVF International Conference on Computer Vision*, pages 25506–25518, 2025.

[6] Rong Li et al. SeeGround: See and ground for zero-shot open-vocabulary 3D visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3707–3717, 2025.

[7] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, et al. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025. OpenReview archive; see also arXiv:2509.07996.

[8] Xuzhi Wang et al. NUC-Net: Non-uniform cylindrical partition network for efficient LiDAR semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):9090–9104, 2025.

[9] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. RoboBEV: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

[10] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised LiDAR semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21705–21715, June 2023.

[11] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn:

Point-voxel feature set abstraction for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[12] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131:531–551, 2023. 2

[13] Ao Liang, Youquan Liu, Yu Yang, Dongyue Lu, Linfeng Li, Lingdong Kong, Huaici Zhao, and Wei Tsang Ooi. LiDARCrafter: Dynamic 4D world modeling from lidar sequences. *arXiv preprint arXiv:2508.03692*, 2025.

[14] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. 1

[15] Till Beemelmanns, Quan Zhang, and Lutz Eckstein. Multicuropt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 3255–3261, 2024. 1

[16] Xiaoshuai Hao et al. SafeMap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025.

[17] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, et al. Is your HD map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, volume 37, pages 22441–22482, 2024. 1

[18] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D multi-object tracking: A baseline and new evaluation metrics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1, 2

[19] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision (ECCV)*, 2020.

[20] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5

[21] John Amanatides and Andrew Woo. A fast voxel traversal algorithm for ray tracing. In *Eurographics*, pages 3–10, 1987.

[22] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. RangeViT: Towards vision transformers for 3D semantic segmentation in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5240–5250, 2023.

[23] Huixian Cheng, Xianfeng Han, and Guoqiang Xiao. CENet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[24] Xiang Xu et al. FRNet: Frustum-range networks for scalable LiDAR segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025. 1

[25] Xinyu Cai, Wentao Jiang, Runsheng Xu, Wenguan Zhao, Jiaqi Ma, Si Liu, and Yikang Li. Analyzing infrastructure LiDAR placement with realistic LiDAR simulation library. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5581–5587, 2023. 1

[26] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3075–3084, 2019. 1, 2, 5

[27] Youquan Liu et al. Multi-space alignments towards universal LiDAR segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14661, 2024.

[28] Lingdong Kong, Xiang Xu, Jiawei Ren, et al. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025. 1

[29] Claudine Badue, Ranãk Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B. Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Bertolini, Thiago M. Paixão, Filipe Mutz, Lucas de Paula Veronese, Thiago Oliveira-Santos, and Alberto F. De Souza. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021. 1

[30] Ye Li et al. Optimizing LiDAR placements for robust driving perception in adverse conditions. *arXiv preprint arXiv:2403.17009*, 2024.

[31] Youquan Liu et al. UniSeg: A unified multi-modal LiDAR segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21662–21673, 2023.

[32] Hengwei Bian et al. DynamicCity: Large-scale 4D occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

[33] Xidong Peng, Runnan Chen, Feng Qiao, et al. Learning to adapt SAM for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024.

[34] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, et al. 3EED: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025. 1

[35] Qi Chen, Sourabh Vora, and Oscar Beijbom. PolarStream: Streaming LiDAR object detection and segmentation with polar pillars. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 26871–26883, 2021. 1

[36] Xiang Xu et al. 4D contrastive superflows are dense 3D representation learners. In *European Conference on Computer Vision (ECCV)*, pages 58–80, 2024.

[37] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3D and 4D panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3480–3495, 2024.

[38] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards robust and reliable 3D perception against corruptions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19994–20006, 2023.

[39] Xiaoshuai Hao, Guanqun Liu, Yuting Zhao, et al. MSC-Bench: Benchmarking and analyzing multi-sensor corruption

for driving perception. *arXiv preprint arXiv:2501.01037*, 2025.

[40] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Qingyuan Zhou, Rui Zhang, Zhijun Li, Wen-Ming Chen, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3D semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):20287–20301, 2025.

[41] Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, and Qingshan Liu. LiMoE: Mixture of LiDAR representation learners from automotive scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27368–27379, 2025.

[42] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. RangeNet++: Fast and accurate LiDAR semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4213–4220, 2019. 1

[43] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

[44] Xuzhi Wang, Xinran Wu, Song Wang, et al. Monocular semantic scene completion via masked recurrent networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24811–24822, 2025.

[45] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for LiDAR segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 228–240, 2023. 1

[46] Lingdong Kong, Xiang Xu, Jun Cen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Calib3D: Calibrating model preferences for reliable 3D scene understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1965–1978, 2025. 2

[47] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Voxel-based 3D detector; applied to autonomous-driving 3D detection on KITTI. 2

[48] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. Voxel-based with sparse 3D conv; AD 3D detection on KITTI/nuScenes. 2, 5

[49] Jiaqi Deng, Shaoshuai Shi, Pei Li, Zhe Zhou, Yizhou Zhang, Hongsheng Li, and Xiaojuan Qi. Voxel R-CNN: Towards high performance voxel-based 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. Voxel-based two-stage detector; AD 3D detection on KITTI/Waymo.

[50] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12605–12614, 2020.

[51] Lingdong Kong, Xiang Xu, Youquan Liu, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025. 2

[52] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[53] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 5

[54] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Point-based two-stage detector; AD 3D detection on KITTI. 2

[55] Weijing Shi and Ragunathan (Raj) Rajkumar. Point-GNN: Graph neural network for 3D object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Point-based with graph neural networks; AD 3D detection on KITTI.

[56] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. 3DSSD: Point-based 3D single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Point-based single-stage; efficient AD 3D detection on KITTI/nuScenes. 2

[57] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your LiDAR placement optimized for 3D scene understanding? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 34980–35017, 2024. 2, 4

[58] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 2

[59] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*. PMLR, 2020. 2

[60] Jungho Kim Junho Koh, Youngwoo Lee. Onlinebev: Recurrent temporal fusion in bird's-eye view representations for multi-camera 3d perception. *arXiv preprint arXiv:2507.08644*, 2025. 2, 4

[61] Lingdong Kong, Niamul Quader, and Venice Erin Liong. ConDA: Unsupervised domain adaptation for LiDAR segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023. 2

[62] Youquan Liu et al. La La LiDAR: Large-scale layout generation from LiDAR data. *arXiv preprint arXiv:2508.03691*, 2025. 2

[63] Dekai Zhu, Yixuan Hu, Youquan Liu, et al. Spiral: Semantic-aware progressive LiDAR scene generation and understanding. *arXiv preprint arXiv:2505.22643*, 2025. 2

[64] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 5

[65] Lingdong Kong, Shaoyuan Xie, Zeying Gong, Ye Li, Meng Chu, Ao Liang, Yuhao Dong, Tianshuai Hu, Ronghe Qiu, Rong Li, Hanjiang Hu, Dongyue Lu, Wei Yin, Wenhao Ding, Linfeng Li, Hang Song, Wenwei Zhang, Yuexin Ma, Junwei Liang, Zhedong Zheng, Lai Xing Ng, Benoit R. Cottereau, Wei Tsang Ooi, Ziwei Liu, Zhanpeng Zhang, Weichao Qiu, Wei Zhang, Ji Ao, Jiangpeng Zheng, Siyu Wang, Guang Yang, Zihao Zhang, Yu Zhong, Enzhu Gao, Xinhan Zheng, Xueting Wang, Shouming Li, Yunkai Gao, Siming Lan, Mingfei Han, Xing Hu, Dusan Malic, Christian Fruhwirth-Reisinger, Alexander Prutsch, Wei Lin, Samuel Schulter, Horst Possegger, Linfeng Li, Jian Zhao, Zepeng Yang, Yuhang Song, Bojun Lin, Tianle Zhang, Yuchen Yuan, Chi Zhang, Xuelong Li, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, Aodi Wu, Xubo Luo, Erjia Xiao, Lingfeng Zhang, Yingbo Tang, Hao Cheng, Renjing Xu, Wenbo Ding, Lei Zhou, Long Chen, Hangjun Ye, Xiaoshuai Hao, Shuangzhi Li, Junlong Shen, Xingyu Li, Hao Ruan, Jinliang Lin, Zhiming Luo, Yu Zang, Cheng Wang, Hanshi Wang, Xijie Gong, Yixiang Yang, Qianli Ma, Zhipeng Zhang, Wenxiang Shi, Jingmeng Zhou, Weijun Zeng, Kexin Xu, Yuchen Zhang, Haoxiang Fu, Ruibin Hu, Yanbiao Ma, Xiyan Feng, Wenbo Zhang, Lu Zhang, Yunzhi Zhuge, Huchuan Lu, You He, Seungjun Yu, Junsung Park, Youngsun Lim, Hyunjung Shim, Faduo Liang, Zihang Wang, Yiming Peng, Guanyu Zong, Xu Li, Binghao Wang, Hao Wei, Yongxin Ma, Yunke Shi, Shuaipeng Liu, Dong Kong, Yongchun Lin, Huitong Yang, Liang Lei, Haoang Li, Xinliang Zhang, Zhiyong Wang, Xiaofeng Wang, Yuxia Fu, Yadan Luo, Djamahl Etchegaray, Yang Li, Congfei Li, Yuxiang Sun, Wenkai Zhu, Wang Xu, Linru Li, Longjie Liao, Jun Yan, Benwu Wang, Xueliang Ren, Xiaoyu Yue, Jixian Zheng, Jinfeng Wu, Shurui Qin, Wei Cong, and Yao He. The RoboSense challenge: Sense anything, navigate anywhere, adapt across platforms. https://robosense2025.github.io, 2025. 4

[66] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, et al. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023. 4

[67] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 21298–21342, 2023. 4

[68] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024. 4

[69] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025. 4

[70] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025. 4

[71] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9122–9129, 2025.

[72] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231, 2024.

[73] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27725–27738, 2025. 4

[74] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810, 2021. 5

[75] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5