DataEngine: Unified Pre-Training and Viewpoint Normalization for Cross-Platform 3D Object Detection

Youngseok Kim[*] Sihwan Hwang[*†] Hyeonjun Jeong[*†]
[*]Visionary Inc. [†]KAIST
{youngseok.kim, sihwan.hwang, hyeonjun.jeong}@visionary.run

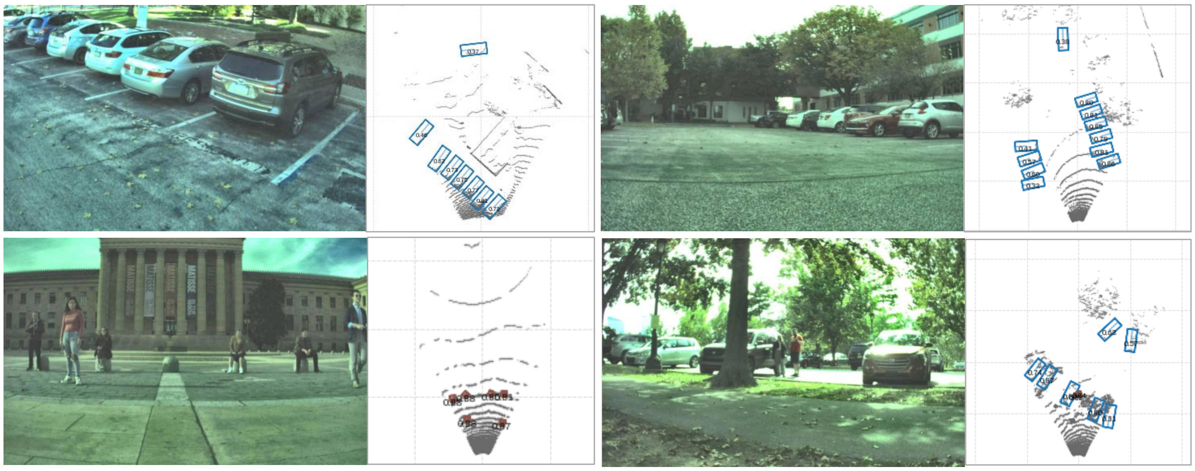

Figure 1. Qualitative results on the RoboSense Cross-Platform Dataset.

Abstract

This technical report provides our first-place solution for RoboSense Challenge - Cross-Platform 3D Object Detection. The primary objective of this challenge is to effectively adapt a model trained on data collected from a vehicle platform to heterogeneous platforms such as drones and quadruped robots. To address this, we employed three core strategies. First, we establish a highly generalizable foundation by pre-training a single model on a unified dataset amalgamated from large-scale public datasets, including Waymo and nuScenes. This approach ensured initial robustness across diverse sensor configurations and environments. Second, to resolve the geometric discrepancies arising from viewpoint differences between platforms, we propose a RANSAC-based ground normalization pre-processing. This canonical representation allows the model to learn viewpoint-invariant features. Finally, we applied Test-Time Augmentation (TTA) to maximize robustness during inference. Our method achieved a winning mAP of 66.94 in phase 1 and 58.54 in phase 2, demonstrating its effectiveness in cross-platform 3D object detection.

1. Introduction

With recent advancements in robotics, 3D object detection utilizing LiDAR sensors has emerged as a core technology for a wide array of autonomous systems [1–9], extending beyond the domain of autonomous driving [10–15].

However, the majority of existing research and datasets have focused on vehicle platforms, assuming a sensor viewpoint largely parallel to the ground [16–21]. Consequently, directly deploying models trained on such data directly onto new platforms, such as drones with an aerial perspective or quadrupeds on uneven terrain, leads to a severe degradation in performance [10]. This domain gap between platforms is primarily caused by geometric inconsistencies in the data distribution, which stem from variations in sensor height, angle, and motion characteristics [22–30].

This challenge presents a critical and practical problem of enabling model adaptation across heterogeneous robotic platforms. Our approach was centered on developing a universal 3D detection model designed for generalization, thereby avoiding overfitting to the characteristics of any single platform.

Our primary contributions are summarized as follows:

- **Unified Large-Scale Pre-training:** To build a robust and

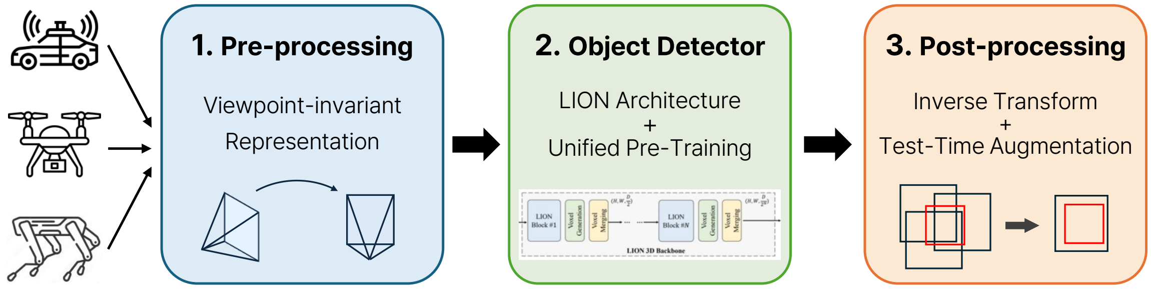

Figure 2. An overview of our proposed methodology, which consists of three stages: (1) Pre-processing: Input data from diverse platforms is canonicalized into a viewpoint-invariant representation. (2) Object Detector: The normalized point cloud is processed by the generalized LION detector. (3) Post-processing: Predictions are mapped back to the original coordinates and fused using Test-Time Augmentation.

  generalizable foundation model, we construct a comprehensive pre-training dataset by merging heterogeneous and large-scale 3D object detection benchmarks. The resulting model serves as a powerful foundation, exhibiting inherent robustness to a wide array of sensor specifications and operational environments without requiring extensive fine-tuning.

- **Viewpoint Normalization Pre-processing:** To address the geometric domain gap arising from disparate sensor viewpoints, we propose a RANSAC-based pre-processing step that canonicalizes all point clouds to a consistent ground-aligned frame. By using RANSAC to estimate the ground plane, we normalize all incoming point clouds into a canonical, ground-aligned frame. This forces the model to learn consistent geometric features, irrespective of the platform's unique perspective.

2. Methodology

Our methodology is structured into three main components as illustrated in Figure 2. First, a pre-processing stage performs viewpoint normalization to align data from all platforms into a canonical frame. Second, our LION-based object detector is trained using a two-phase strategy: comprehensive pre-training on a unified large-scale dataset followed by fine-tuning on the official challenge data. Finally, a post-processing stage uses Test-Time Augmentation (TTA) during inference to maximize performance.

2.1. Baseline Model

We selected the LION [31] architecture as our baseline model. Unlike prevalent approaches that rely on Sparse Convolution (spconv) [32], LION is a window-based framework built upon modern state-space models (SSMs), such as RetNet [33] and Mamba [34]. The primary advantage of Linear RNNs is their linear computational complexity with respect to sequence length, a significant improvement over the quadratic cost of transformers. The sequential processing nature of SSMs allows them to capture long-range spatial dependencies more effectively than CNN-based methods, leading to a larger effective receptive field and superior generalization performance across diverse point cloud distributions.

Feature Extractor: The LION feature extractor is a hierarchical 3D backbone composed of several LION Blocks. Each block synergistically models long-range dependencies using a Linear Group RNN and captures local geometric details with a 3D sub-manifold convolution-based descriptor. This descriptor is critical for re-injecting spatial information lost when 3D data is flattened into sequences for the RNN. Additionally, the model leverages the auto-regressive properties of the RNN to implement a voxel generation strategy, which densifies sparse foreground features. This hybrid approach allows the encoder to generate rich, multi-scale feature representations that effectively capture both the global context and local structure of the 3D scene.

Detection Head: The features from the encoder are consumed by a detection head to produce 3D bounding box predictions. We employ a CenterPoint [35] style detection head, which first predicts object centers via a heatmap and then regresses other object attributes, such as location, dimensions, and orientation, from the features at these center locations.

Loss Function: The total loss L_{total} is a weighted sum of the classification loss and the regression loss, defined as:

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg}$$

where λ_{cls} and λ_{reg} are the corresponding weighting factors. We employ the Focal Loss [36] for classification loss L_{cls} to address class imbalance and a smooth L1 Loss for regression loss L_{reg} to predict the bounding box parameters, including its center offset (x, y, z), dimensions (w, l, h), and orientation (θ).

2.2. Unified Large-Scale Pre-training

To build a foundational model robust to diverse platforms and environments, we adopted a strategy of pre-training a single model on a unified corpus of six large-scale public datasets: Waymo [2], nuScenes [1], ONCE [37], Pandaset [5], Lyft [38], and Argoverse2 [3]. Each of these datasets is characterized by different LiDAR sensors, sensor mounting configurations, and geographical environments. By integrating this vast and varied data, the model learns a universal 3D object representation that transcends the biases of any single dataset.

However, unifying multiple datasets entails technical challenges beyond simple data concatenation, as each feature has distinct class definitions, detection ranges, and annotation formats. We addressed these inconsistencies as follows:

Unified Class Mapping: We standardized class definitions across all datasets to establish a consistent learning target. For example, we reconciled the coarse-grained classes from Waymo [2] (*e.g.*, Car, Pedestrian, Cyclist) with the 20+ fine-grained classes from Argoverse2 [3] by designing a hierarchical class structure that the model learns to predict.

Range-Adaptive Loss Masking: To account for the varying annotation ranges across datasets, we applied a loss masking technique. This method excludes regions outside the annotated range of a specific dataset from the loss calculation, preventing the model from being penalized for predictions in unannotated areas.

2.3. Viewpoint Normalization

A core challenge arises from the geometric shift between platforms: vehicle-mounted LiDARs have a ground-parallel viewpoint, drones have an aerial view, and quadrupeds have an unstable, terrain-dependent perspective. This variation causes identical objects to be represented with vastly different point cloud signatures, introducing significant ambiguity for the detection model.

To address this challenge, we implemented a preprocessing step to align all point clouds to a consistent ground-relative coordinate frame. First, we use the RANSAC algorithm [39] to robustly estimate the ground plane's normal vector, n. We then compute a rigid-body transformation, defined by a rotation R and translation t, that aligns this plane with the standard XY-plane. Every point p in the input cloud is mapped to a normalized point p' using this transformation:

$$p' = Rp + t$$

This process provides the model with a consistent, viewpoint-invariant representation of the scene.

After the model predicts bounding boxes $\{B'\}$ in the normalized space, we apply the inverse transformation to map them back to the original coordinate system. The center c' and orientation θ' of a predicted box are recovered in the original frame by inverse transformation. This ensures our final predictions are correctly oriented and positioned for evaluation.

2.4. Test-Time Augmentation (TTA)

To maximize performance during the final inference stage, we employed a Test-Time Augmentation (TTA) strategy. For each test sample, we generated an ensemble of augmented versions and averaged their predictions. The applied augmentations included:

- **Flip:** Mirroring the point cloud along the Y-axis.
- **Rotation:** Applying rotations around the Z-axis with angles from the set $\{-20°, -10°, 10°, 20°\}$.

The 3D Bounding box predictions from each augmented view were transformed back to the original coordinate system. We then used Weighted Box Fusion (WBF) [40] to merge the multiple sets of predictions into a single, high-confidence output.

3. Experiments

3.1. Experimental Setup

Datasets: We use the official data provided by the *RoboSense Challenge 2025* [41] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [42, 43] at ICRA 2023 and the *RoboDrive Challenge 2024* [44, 45] 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 [10, 46–51]. Specifically, this task is built upon the **Pi3DET** benchmark [10] in **Track 5**, which studies cross-platform LiDAR-based 3D object detection across vehicle, drone, and quadruped platforms through viewpoint normalization and unified pre-training.

The dataset for this challenge comprises LiDAR point clouds and corresponding 3D bounding box annotations for 'Car' and 'Pedestrian' classes, collected from three distinct platforms: a ground vehicle, a drone, and a quadruped robot. The provided point clouds are pre-filtered, containing only the points that fall within the field of view of a synchronized camera. The dataset is partitioned into 10,032 training samples, 5,885 validation samples for phase 1, and 8,025 validation samples for phase 2.

Evaluation Metric: The official challenge metric is the mean Average Precision (mAP). This is calculated based on the 3D Intersection over Union (IoU) between predicted and ground-truth boxes, using a fixed IoU threshold of 0.5.

Implementation Details: Our implementation is built upon the LION [31] architecture, utilizing RetNet [33] as the feature extractor with feature dimension of 64. The unified pre-training phase was conducted for 6 epochs across the six public datasets. Subsequently, the model was fine-tuned on the official challenge training set for 2 epochs.

For both stages, we used the AdamW optimizer [52] with a cosine annealing learning rate policy, setting the initial learning rate to 2×10^{-4} and weight decay to 0.05. Pre-training was performed on 8 NVIDIA A100 GPUs, while fine-tuning and inference were run on 4 NVIDIA RTX 3090 GPUs.

3.2. Results

Final Leaderboard Results: Our proposed framework achieved first place on the final leaderboards for both phase 1 and phase 2 of the competition. As shown in Table 1, our method obtains a final mAP of 66.94 in phase 1 and 58.54 in phase 2, demonstrating its strong adaptation capabilities across all three platforms.

Table 1. Final results on the official challenge leaderboards.

Phase	Car	Ped.	Avg.
1	66.94	-	**66.94**
2	64.17	52.92	**58.54**

Ablation Study: As detailed in Table 2, we began with a unified pre-trained model as a baseline and progressively added our proposed techniques: fine-tuning on challenge data and Viewpoint Normalization. The results confirm that each component provides a significant performance gain. Note that TTA was applied to all experimental runs to ensure a fair comparison of the core model enhancements.

Table 2. Ablation study on phase2 validation set.

Method	Car	Ped.	Avg.	Δ
Pre-training	59.16	47.73	53.45	-
+ Fine-tuning	62.58	50.30	56.44	+2.99
+ Viewpoint	64.17	52.92	58.54	+5.09

Qualitative Analysis: Figure 1 provides qualitative detection results on challenging scenes from the drone and quadruped platforms. The visualizations confirm our model's ability to accurately detect objects despite extreme viewpoint shifts and occlusion scenarios where conventional vehicle-centric models typically fail. This robust performance underscores the effectiveness of our viewpoint normalization and large-scale pre-training strategies in bridging the cross-platform domain gap.

4. Conclusion

In this technical report, we presented our first-place solution for enhancing 3D object detection performance across heterogeneous robotic platforms. By combining two core strategies: unified large-scale pre-training for a highly generalizable foundation, and viewpoint normalization to resolve geometric distortions, our method demonstrated a high degree of adaptability. We hope our findings can contribute to the advancement of 3D perception for a broader range of robotic applications beyond autonomous vehicles.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[3] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

[4] 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*, pages 3354–3361, 2012.

[5] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. PandaSet: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*. IEEE, 2021.

[6] Ao Liang et al. WorldLens: Full-spectrum evaluations of driving world models in real world. *arXiv preprint arXiv:2512.10958*, 2025.

[7] Lingdong Kong et al. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

[8] Tianyi Yan, Tao Tang, Xingtai Gui, Yongkang Li, Jiasen Zhesng, Weiyao Huang, et al. AD-R1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models. *arXiv preprint arXiv:2511.20325*, 2025.

[9] Tianshuai Hu et al. Vision-language-action models for autonomous driving: Past, present, and future. *arXiv preprint arXiv:2512.16760*, 2025.

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

[11] Rong Li et al. 3EED: Ground everything everywhere in 3D. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[12] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ED: Multi-robot, multi-sensor, multi-environment event dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4023, 2023.

[13] 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*, pages 3707–3717, 2025.

[14] Lingdong Kong, Dongyue Lu, Xiang Xu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. EventFly: Event camera perception from ground to the sky. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1484, 2025.

[15] Song Wang et al. Forging spatial intelligence: A roadmap of multi-modal data pre-training for autonomous systems. *arXiv preprint arXiv:2512.24385*, 2025.

[16] 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*, pages 12605–12614, 2020.

[17] 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.

[18] Xiang Xu et al. LiMoE: Mixture of LiDAR representation learners from automotive scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27368–27379, 2025.

[19] 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.

[20] 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*, pages 228–240, 2023.

[21] 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.

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

[23] Jingyi Xu, Weidong Yang, et al. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):20287–20301, 2025.

[24] 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.

[25] 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.

[26] Lingdong Kong et al. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025.

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

[28] Ao Liang et al. LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. *arXiv preprint arXiv:2508.03692*, 2025.

[29] Dekai Zhu et al. Spiral: Semantic-aware progressive LiDAR scene generation and understanding. *arXiv preprint arXiv:2505.22643*, 2025.

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

[31] Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. Lion: Linear group rnn for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 2024.

[32] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022.

[33] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

[34] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

[35] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[36] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[37] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021.

[38] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*. PMLR, 2021.

[39] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, (6), 1981.

[40] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 2021.

[41] 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.

[42] 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, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[43] Lingdong Kong et al. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, volume 36, pages 21298–21342, 2023.

[44] 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.

[45] Shaoyuan Xie et al. 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.

[46] Shaoyuan Xie et al. 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.

[47] 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*, pages 9122–9129, 2025.

[48] Ye Li et al. Is your LiDAR placement optimized for 3D scene understanding? In *Advances in Neural Information Processing Systems*, volume 37, pages 34980–35017, 2024.

[49] 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.

[50] 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*, pages 19994–20006, 2023.

[51] Shaoyuan Xie et al. RoboBEV: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.