Enhancing Robust Social Navigation with Cutout: Handling Human Occlusion in Dynamic Environments

Faduo Liang
South China University of Technology
Guangzhou 510641, China
aulfd@mail.scut.edu.cn

Abstract

This report presents a simple yet effective approach to enhancing social navigation in dynamic human environments, built upon the Falcon framework. While Falcon achieves socially aware navigation by explicitly forecasting human trajectories and penalizing actions that interfere with future human paths, it remains vulnerable to occlusion-induced failures when operating with limited-viewpoint sensors such as RGB-D cameras. These occlusions frequently occur when humans suddenly appear from lateral or rear angles, reducing the reliability of learned navigation policies. To address this limitation, we introduce an image-based data augmentation strategy that simulates occlusion during training. Specifically, we inject random black patches into the depth observations, emulating partial visibility conditions and encouraging the model to learn more conservative and robust navigation behaviors. Experimental results on the RoboSense 2025 Track 2 benchmark show that this simple augmentation substantially improves navigation stability and safety under restricted viewpoints, demonstrating its practical value for real-world social navigation.

1. Introduction

Social navigation, which requires robots to move safely and naturally among humans, has been an active research area for decades [1–3]. Robots operating in tight indoor environments must not only avoid collisions but also respect social norms and human motion patterns, calling for advanced RGB-D perception and robust decision-making systems [4–11].

Early navigation approaches primarily focused on static or quasi-static environments [12]. More recent work incorporates human motion understanding, including intent prediction, risk estimation, and trajectory forecasting, to enable anticipatory behaviors [13–15]. Falcon [1], for instance, explicitly predicts future human trajectories and penalizes navigation actions that obstruct likely human paths, achieving notable improvements in social compliance.

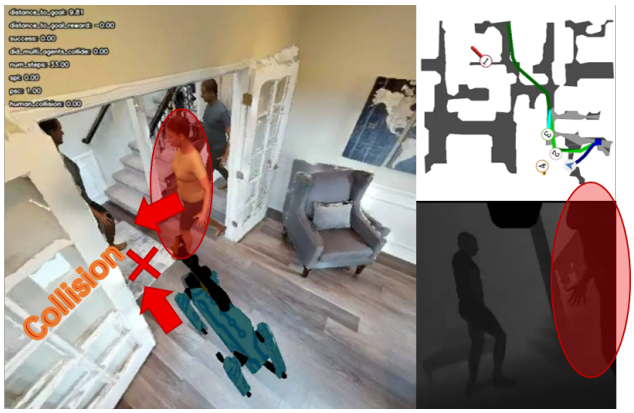


Figure 1. Collision occurs when a pedestrian suddenly approaches the robot from the side and obstructs its view.

Despite these advancements, a persistent challenge remains: limited sensor viewpoints. Robots equipped with RGB-D sensors often experience occlusion when people appear suddenly from the sides or rear, causing an abrupt loss of visibility [1, 16]. Under these conditions, models that rely heavily on line-of-sight observations may fail to anticipate human motion, leading to unsafe or unstable navigation behavior (see Figure 1).

Data augmentation has proven effective in improving visual robustness in different learning settings [12, 17–21]. By exposing agents to perturbed or partially corrupted observations, augmentation can encourage policies to generalize beyond idealized sensor inputs [22–25].

Motivated by these insights, we investigate whether simple image-based augmentation can enhance robustness in social navigation. Our method applies random black patches to depth maps during training, mimicking occluded or partially obstructed views. Without modifying the Falcon architecture or introducing additional auxiliary tasks, this lightweight strategy significantly improves navigation performance un-

der occlusion-heavy scenarios.

Our findings highlight that even minimal augmentation can substantially strengthen robustness in RGB-D-based navigation, offering a practical and easily deployable enhancement for real-world social robotics systems.

2. Related Work

2.1. Social Navigation

Social navigation aims to enable robots to move safely and naturally in human-populated environments, requiring awareness of human motion patterns, social norms, and implicit interaction cues. The DD-PPO framework [12] serves as a foundational reinforcement learning paradigm for this task and has inspired numerous extensions that incorporate human-centric reasoning. For example, Proximity-Aware Tasks introduced in [13] augment navigation policies with auxiliary objectives that encourage robots to infer social cues related to risk and interpersonal distance. More recently, Falcon [1] proposes a socially compliant navigation strategy by explicitly predicting future human trajectories and penalizing robot actions that obstruct likely human paths.

Building on these efforts, our work adopts the Falcon framework as the base navigation model and leverages the Habitat simulation platform for controlled training and evaluation. The agent operates using egocentric depth observations and egomotion inputs, reflecting a realistic onboard sensing configuration for indoor mobile robots.

2.2. Data Augmentation in Reinforcement Learning

Data augmentation has long been recognized as a powerful technique for improving generalization in computer vision [26–31]. Its importance has subsequently expanded into reinforcement learning (RL), where training data is often limited, and policies must generalize across diverse visual conditions. The study in [17] provides a comprehensive analysis of augmentation strategies for both pixel-based and state-based RL, demonstrating that simple transformations can significantly enhance robustness and sample efficiency. In dynamic environments, combining visual augmentation with environment-driven variability can further strengthen generalization, as shown in [32].

Motivated by these findings, our work explores whether a lightweight image-based augmentation method can improve robustness in social navigation. Specifically, we incorporate a traditional occlusion-mimicking augmentation into the Falcon framework [1], enabling the navigation policy to better handle sudden visibility loss caused by real-world occlusion scenarios.

3. Methodology

Image-based data augmentation has proven highly effective in improving the robustness of reinforcement learning (RL) policies, especially when training relies on high-dimensional visual observations. Prior work [17] demonstrates that simple augmentations – such as random translations, crops, color jittering, random convolutions, amplitude scaling, and patch cutout – can enable lightweight RL algorithms to match or surpass more complex state-of-the-art approaches across a variety of pixel-based benchmarks. Among these techniques, *cutout* has emerged as particularly valuable for modeling object occlusions, a common challenge in tasks such as object recognition, tracking, and human pose estimation [33]. It is also well aligned with convolutional architectures, including the ResNet-50 backbone adopted by Falcon.

Motivated by these observations, we employ the cutout augmentation strategy as our primary mechanism for improving robustness in social navigation. Following the formulation in [17, 18], cutout operates by inserting a black rectangular patch into the input depth image (Figure 2). This patch is assigned a zero-depth value and is randomly sampled in size, aspect ratio, and spatial location, effectively simulating partial occlusions that may arise when humans suddenly enter the robot's field of view from unobserved directions.

We adopt the parameterization recommended by Zhong et al. [18]: the patch aspect ratio is sampled from the range $[0.3, 3.33]$, and the relative area is sampled from $[0.02, 0.33]$ of the original frame. These settings introduce sufficient variability to mimic realistic occlusion patterns without overly corrupting the sensory input. Importantly, this augmentation is applied at the input level and does not require any modification to the Falcon architecture, training pipeline, or auxiliary objectives.

By integrating cutout into the training process, we aim to expose the navigation policy to a broader distribution of partially observable states. This encourages the agent to adopt more conservative and socially compliant behaviors when visibility is limited, ultimately improving robustness in real-world scenarios where occlusions are frequent and unpredictable.

4. Experiments

In this section, we first introduce the benchmark and its two datasets, followed by the experimental setup and evaluation metrics. Next, we present the results of our method compared to prior approaches. Finally, we analyze why our method works.

4.1. Experimental Setups

Dataset. We use the official data provided by the *RoboSense Challenge 2025* [34] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [35, 36] at ICRA 2023 and the *RoboDrive Challenge 2024* [37, 38] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in

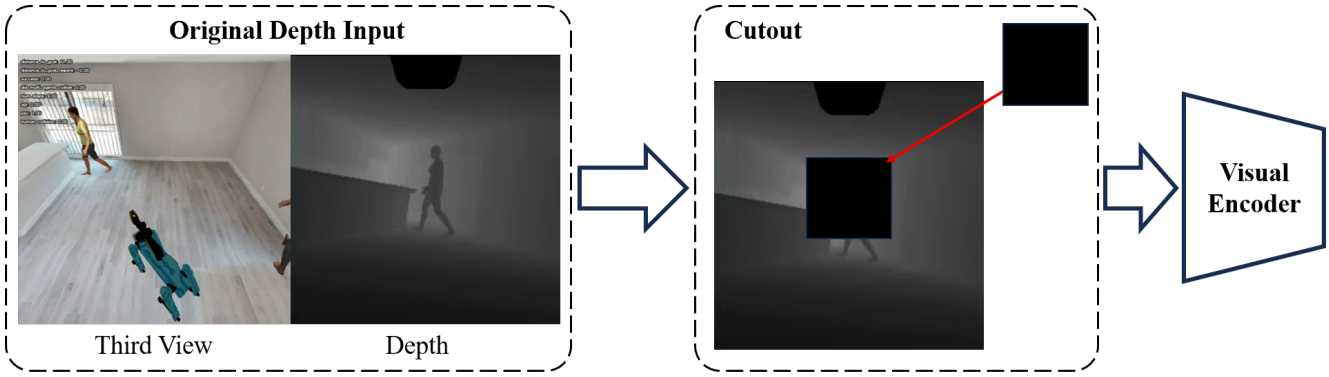


Figure 2. Image-based data augmentation from [17]. We use CutOut to preprocess the original images before feeding them into the ResNet-50 network for visual encoding.

Table 1. Performance Evaluation of SocialNav Tasks on Social-HM3D (upper group) and RoboSense-Track2-Social-Navigation-Dataset (lower group). Data in the table represents percentages. We **bold** the best results and underline the second best results. ↓ means the lower the metric, the better the performance, and vice versa. Bold numbers represent the metric-wise best performance.

Dataset	Method	Suc. ↑	SPL ↑	PSC ↑	H-Coll ↓	Total ↑
Social-HM3D	Falcon	55.15	**55.15**	89.56	42.96	65.47
	Ours	**57.77**	51.71	**89.78**	**39.65**	**65.55**
RoboSense-Track2	Falcon	54.00	49.97	**86.30**	39.20	62.48
	Ours	**65.20**	**58.55**	86.11	**32.60**	**69.48**

this competition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [1, 39–42].

Metrics. Our benchmark metrics build upon existing work [1] and focus on two principal perspectives: task completion and adherence to SocialNav objectives. For task completion, we use Success Rate (Suc.) and Success weighted by Path Length (SPL). For social norms, we evaluate Human-Robot Collision Rate (H-coll) and Personal Space Compliance (PSC). Considering the human collision radius is 0.3m and the robot is 0.25m, the PSC distance threshold is set to 1.0m. The final score is computed as:

$$Total = 0.4 \times Suc + 0.3 \times SPL + 0.3 \times PSC \ ,$$

to evaluate the overall effectiveness of navigation.

Baseline Models. We primarily compare our method with a recent state-of-the-art social navigation approach, namely Falcon [1]. It introduces the Social Cognition Penalty (SCP), a set of penalties designed to promote adherence to social norms, which leverages three auxiliary tasks to model the number, position, and future trajectory of humans, thereby effectively capturing their current proximity. All methods use only depth images as visual inputs to ensure a fair comparison.

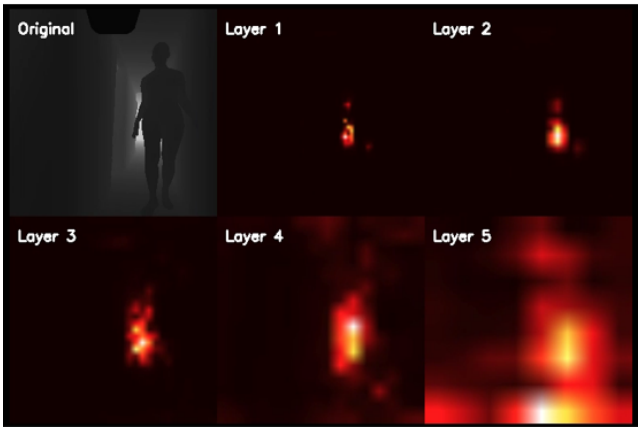
Implementation Details. We train the RL agents using the DD-PPO algorithm [12] with identical hyperparameters. Our model is initialized with pre-trained weights from a PointNav model [43] and fine-tuned for 10 million steps on the SocialNav task. Training is conducted on 4 Nvidia RTX 3090 GPUs with 8 parallel environments. Models are trained on the Social-HM3D train set and tested on both Social-HM3D and the RoboSense Track2 social navigation dataset. The checkpoint yielding the highest score is selected.

4.2. Result Analysis

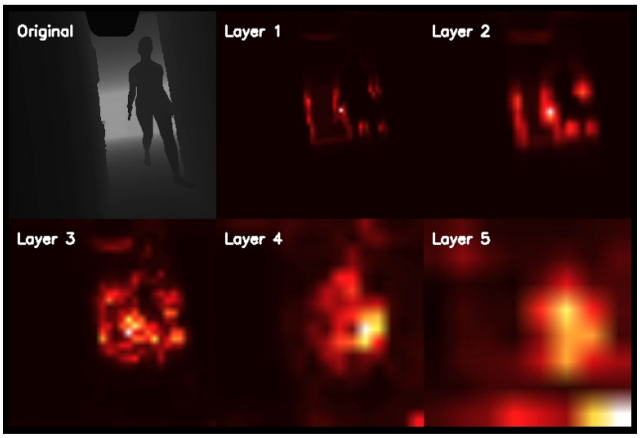
Table 1 presents the comparison results on the SocialHM3D test set and the RoboSense Track2 social navigation dataset. Asshown, the superior results demonstrate the effectiveness of our method in both goal-reaching and social compliance.

The result indicate that our method effectively increases the success rate, achieving 2.62% and 11.20% improvements, respectively, and reduces the risk of human collisions, with 3.11% and 6.60% reductions, respectively, while maintaining PSC indicators largely consistent. Fluctuations in SPL suggest that our approach avoids collisions while pursuing shorter paths, striking a better balance between navigation efficiency and safety.

These results validate the ideas from [17] and [32] that reinforcement learning with augmented data is simple but effective, and that domain-specific data augmentation can



(a) Spatial attention map for the baseline Falcon model.



(b) Spatial attention map for our Cutout-augmented model.

Figure 3. The spatial attention maps from the encoder illustrate the regions on which the agent focuses to make navigation decisions. Compared to the baseline agent trained without augmentation, Cutout enables the agent to concentrate on the human body while ignoring irrelevant scene details.

further increase performance gains when combined with image augmentation methods. Similarly, we compute the spatial attention maps by mean-pooling the absolute values of the activations along the channel dimension, followed by a 2-dimensional spatial softmax applied to five layers of the Resnet-50 network. Figure 3 visualizes the spatial attention maps for the the baseline Falcon model and our Cutout-augmented model. Based on these findings, we illustrate why cutout works:

Cutout contributes to navigation performance improvement by simulating actual occlusion Adjacent pixels in an image share much of the same information. Cutout forces the network to reconstruct the image from the remaining pixels, encouraging it to better utilize the complete image context rather than relying on a small set of specific visual features, thereby achieving a deeper understanding.

Cutout's effect on activations. As shown in figure 3, cutout increases activation strength in the shallow layers of the network, while in deeper layers, more activations appear in the tail of the distribution, which enhances the robustness and overall performance of convolutional neural networks.

5. Discussion & Conclusion

In this work, we introduce a simple yet effective method to enhance social navigation in dynamic human environments by integrating Cutout augmentation into the Falcon framework, simulating occlusions via random black patches on depth inputs to mitigate limitations in prior trajectory prediction models under constrained RGB-D viewpoints. Evaluations on Social-HM3D and RoboSense Track 2 datasets reveal substantial gains, with our model achieving up to 11.20% higher success rates and 6.60% lower human-robot collisions compared to baseline Falcon, while preserving personal space compliance; spatial attention maps further affirm augmentation's reinforcement learning efficacy by demonstrating sharper focus on human features amid reduced scene distractions. These findings corroborate prior studies on augmented RL and underscore Cutout's domain-specific potential to fortify robustness in safety-critical autonomous tasks, offering a lightweight, plug-and-play solution extensible to other occlusion-resilient embodied AI applications.

Despite these advances, our method – primarily a minimal integration of Cutout augmentation into the baseline Falcon framework – retains certain limitations. Evaluations remain confined to simulated environments via the Habitat platform, potentially overlooking real-world factors such as sensor noise, variable lighting, or erratic human dynamics.

Future work could explore real-world deployments on physical robots to validate generalization, investigate adaptive augmentation strategies that dynamically adjust patch parameters based on environmental context, and fuse multi-sensor data for more comprehensive occlusion handling. Ultimately, our work contributes to the advancement of safer autonomous navigation systems, thereby aligning with the objectives of the 2025 RoboSense Challenge.

References

[1] 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.

[2] Angelique M Taylor, Sachiko Matsumoto, Wesley Xiao, and Laurel D Riek. Social navigation for mobile robots in the emergency department. In *IEEE International Conference on Robotics and Automation*, pages 3510–3516, 2021.

[3] Aditya Kapoor, Sushant Swamy, Pilar Bachiller, and Luis J Manso. SocNavGym: a reinforcement learning gym for social navigation. In *IEEE International Conference on Robot and Human Interactive Communication*, pages 2010–2017, 2023.

[4] 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.

[5] Runnan Chen et al. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

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

[7] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 75896–75910, 2023.

[8] Zeying Gong et al. Stairway to success: An online floor-aware zero-shot object-goal navigation framework via LLM-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*, 2025.

[9] 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*, pages 27368–27379, 2025.

[10] 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.

[11] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

[12] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect point-goal navigators from 2.5 billion frames, 2020.

[13] Enrico Cancelli, Tommaso Campari, Luciano Serafini, Angel X. Chang, and Lamberto Ballan. Exploiting proximity-aware tasks for embodied social navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10923–10933, 2023.

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

[15] 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.

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

[17] Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *arXiv preprint arXiv:2004.14990*, 2020.

[18] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation, 2017.

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

[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] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 37193–37229, 2023.

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

[23] 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.

[24] 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.

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

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] 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.

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

[29] 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.

[30] Jingyi Xu 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.

[31] 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.

[32] Naoki Yokoyama, Qian Luo, Dhruv Batra, and Sehoon Ha. Learning robust agents for visual navigation in dynamic environments: The winning entry of igibson challenge 2021. *ArXiv*, abs/2109.10493, 2021.

[33] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[34] 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.

[35] 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.

[36] 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*, volume 36, pages 21298–21342, 2023.

[37] 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, et al. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[38] 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.

[39] 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.

[40] 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*, volume 37, pages 34980–35017, 2024.

[41] 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.

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

[43] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021.