RoboSense:
The Robust Robot Sensing Challenge

Lingdong Kong, Shaoyuan Xie, Ye Li, Xiaohao Xu, Feng Xue,
Meng Chu, Yaru Niu, Hanjiang Hu, Zhedong Zheng,
Xiaonan Huang, Benoit R. Cottereau, Lai Xing Ng,
Wei Tsang Ooi, Wenwei Zhang, Liang Pan, Ziwei Liu

Proposal to IROS 2025 Competition

https://robosense2025.github.io

1 Introduction

The RoboSense Challenge aims to push the boundaries of **robust robot perception in challenging conditions**, addressing the growing need for resilient and adaptive systems across various robotic platforms and environments.

As autonomous systems become increasingly integrated into real-world applications, their ability to maintain high perception accuracy in the face of sensor noise, environmental perturbations, and unpredictable conditions becomes a critical factor in their overall performance [6,10,15]. These challenges are particularly important for tasks like autonomous driving, robot navigation, cross-view localization, and simultaneous localization and mapping (SLAM), where robust perception is crucial for safety and operational reliability [3,14,18].

Background. At last year's ICRA in Yokohama, we successfully organized the **RoboDrive Challenge** [9], focusing on enhancing robust perception in autonomous driving under out-of-distribution conditions, such as sensor failures and adverse weather. The competition targeted four key tasks: BEV detection, map segmentation, semantic occupancy prediction, and multi-view depth estimation, which are critical for reliable navigation and decision-making in dynamic environments. The challenge **attracted 140 teams from 93 institutes across 11 countries**, with nearly 1,000 submissions across five tracks. Participants introduced cutting-edge techniques, including multi-sensor fusion

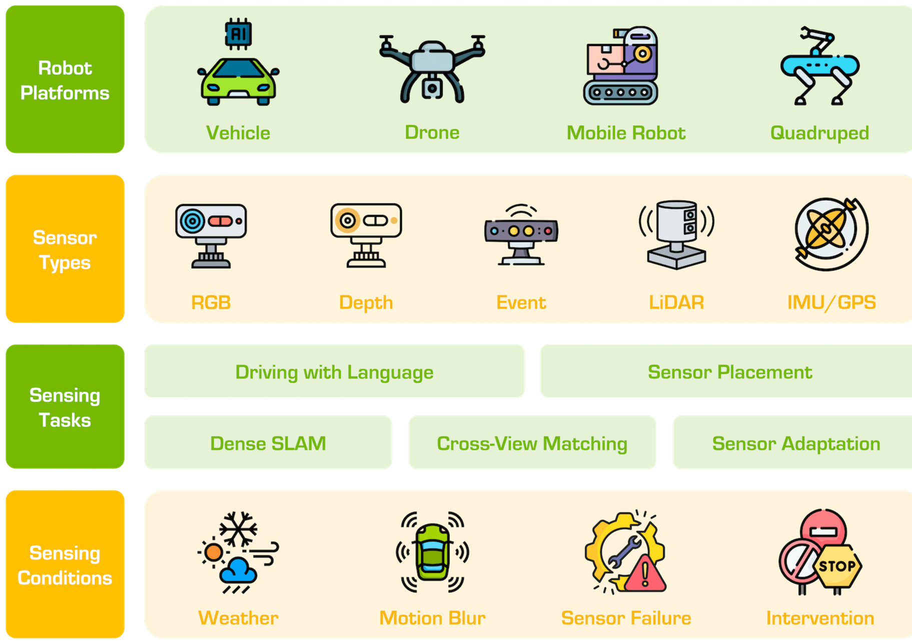

Figure 1: Overview of the robot platforms, sensor types, robust sensing tasks, and various sensing conditions covered in this competition. Our competition covers a wide range of autonomous platforms, including vehicles, drones, and quadrupeds, equipped with diverse sensor types such as RGB cameras, LiDAR, event cameras, and IMU/GPS. The sensing tasks span from robust object detection and mapping to semantic segmentation, SLAM, and cross-view matching, all under challenging conditions such as adverse weather, sensor malfunctions, motion blur, and environmental noise. The goal is to develop models that maintain high accuracy and resilience despite these real-world variabilities, ensuring reliable performance across different robotic platforms and environments.

and advanced data augmentation, to improve perception robustness under challenging conditions. The top solutions demonstrated significant advancements, offering new insights into the resilience of autonomous driving perception systems.

Motivation & Overview. Building on the success of RoboDrive, this new competition extends the challenge to a broader range of robot perception tasks, ensuring robust performance across diverse scenarios. **RoboSense** is designed to challenge participants to develop innovative solutions that enhance the resilience of robot perception models under various challenging conditions, such as sensor failures, adverse weather, motion blur, and data corruption. The competition offers **five distinct tracks**, each addressing a specific aspect of robust robot perception. Figure 1 provides an overview of the robot platforms, sensor types, robust sensing tasks, and various sensing conditions covered in this competition.

- **Track #1: Robust Driving with Language** focuses on vision-language models that fuse perception, prediction, and planning with natural language understanding to improve the robustness of autonomous driving systems under natural corruption scenarios. This track targets real-world conditions where sensor degradation or environmental challenges can impair the system's ability to make accurate driving decisions.

- **Track #2: Robust Sensor Placement** addresses the optimization of sensor placement for autonomous vehicles. Participants are tasked with developing models that can adapt perception systems to different sensor layouts while maintaining high accuracy in adverse conditions, such as weather effects and sensor failures. The goal is to ensure that perception systems remain robust regardless of the placement and configuration of the sensors.

- **Track #3: Robust Dense SLAM** challenges participants to develop dense SLAM systems that generate accurate 3D reconstructions and trajectory estimations under noisy inputs. This track focuses on improving the robustness of SLAM models in environments with sensor noise, lighting variations, and dynamic objects, all of which can compromise mapping and localization.

- **Track #4: Robust Cross-View Matching** focuses on matching visual and textual data across drastically different perspectives, such as aerial (drone or satellite) and ground-level viewpoints. Participants are tasked with developing models that can robustly match corresponding elements across views, even under natural corruptions and environmental changes.

- **Track #5: Robust Sensor Adaptation** challenges participants to design perception models that can adapt seamlessly across different platforms, including vehicles, drones, and quadrupeds. The models must generalize from vehicle-based training to other platforms, using sensor data from event cameras, RGB cameras, LiDAR, and IMU/GPS, while maintaining high perception accuracy across varying sensor configurations and motion patterns.

Relation to IROS. The spirit of the RoboSense Challenge stems from the increasing complexity and diversity of real-world environments in which autonomous systems operate. Whether in autonomous driving, drone navigation, or legged robot exploration, these systems are frequently exposed to adverse conditions that can degrade their sensor data quality. Robust perception is key to ensuring the safe and reliable operation of these systems, making it a critical research area for the IROS community. Furthermore, developing models and algorithms that can generalize across different platforms and sensor setups without requiring extensive retraining will be crucial for scaling autonomous technologies across multiple industrial applications.

Each track in the competition reflects a core challenge in the field of robotics, offering participants the opportunity to develop novel techniques for improving the robustness of perception systems. These efforts will not only lead to more resilient models but also pave the way for safer, more reliable autonomous systems capable of handling the diverse and unpredictable challenges of real-world applications.

2 Rules and Assessments

In this **RoboSense** competition, we plan to host **five tracks** for facilitating the development of robust robot sensing techniques in challenge conditions.

2.1 Track #1: Robust Driving with Language

Theme & Motivation. This track challenges participants to develop vision-language models that enhance the robustness of autonomous driving systems under real-world conditions, including sensor corruption and environmental noise [19]. The goal is to build systems that can generate accurate perception, prediction, and planning outputs for vehicle navigation, even when presented with noisy or degraded input data. Urban environments are inherently complex, with challenges such as adverse weather, occlusions, and varying sensor quality, all of which can impair the reliability of current autonomous driving systems.

Participants are expected to design models that fuse driving perception, prediction, and planning with natural language understanding, enabling the vehicle to make accurate, human-like decisions. The models must demonstrate robustness in handling natural corruption scenarios such as sensor failures, data dropouts, and environmental perturbations. These models will be trained and evaluated using a combination of clean and corrupted datasets, challenging participants to ensure their models perform consistently in both normal and degraded conditions.

This track addresses critical challenges in real-world autonomous driving, focusing on improving the ability of autonomous vehicles to understand and react appropriately to their surroundings, despite adverse conditions. The vision-language models developed in this track will also contribute to more intuitive, human-like decision-making systems in autonomous driving applications.

Datasets. The datasets used in this track include:

- *nuScenes* [1], a comprehensive dataset for autonomous driving, including sensor data from six cameras, one LiDAR, five radars, a GPS/IMU unit, as well as rich annotations for different driving tasks.

- *nuScenes-C* [19], a corruption dataset from the RoboBEV benchmark, which extends nuScenes for evaluating the robustness of driving perception models in corrupted environments. It simulates various adverse conditions such as sensor failures, weather effects (*e.g.*, rain, fog, snow), and noise.

- *DriveLM-nuScenes* [16], a newly introduced dataset that combines nuScenes data with natural language descriptions of driving scenarios, enabling the development of models that understand and respond to language-based driving instructions and queries.

These datasets provide a robust platform for developing vision-language models that must handle a range of inputs, from clean data to highly corrupted or noisy data, simulating real-world driving challenges.

Baseline Model. The baseline model for this track is BLIP-2 [11], a vision-language model designed for various multi-modal tasks. It is capable of integrating visual data with language inputs to generate meaningful predictions and responses. In the context of this track, BLIP-2 will serve as the foundation for developing robust models that can process both BEV perception data and natural language inputs to navigate complex driving scenarios.

Evaluation Metrics. The performance of models in this track will be assessed using a combination of perception, behavior, and motion-related metrics. These metrics will comprehensively evaluate the robustness, accuracy, and decision-making capabilities of the models in autonomous driving scenarios:

- SPICE: Measures the semantic content of generated descriptions.
- GPT-Score: Assesses the relevance and quality of the generated language outputs.
- BLEU, ROUGE-L, METEOR, and CIDEr: Standard metrics for evaluating the quality of text outputs, focusing on language generation tasks such as scene description or action planning.
- Accuracy: Measures how accurately the model interprets and responds to language instructions for driving behavior, such as stop, turn, or navigate.
- ADE (Average Displacement Error): Measures the average error between predicted and actual future positions of ego-vehicle in the scene.
- FDE (Final Displacement Error): Measures the error at the final predicted position of an object relative to its ground truth.
- Collision Rate: Evaluate the ability to predict and plan trajectories that avoid collisions with obstacles, other vehicles, and pedestrians.

The primary focus of the evaluation will be on how well models handle natural corruption scenarios in both perception and planning tasks, ensuring that they can generate robust and accurate outputs under a variety of challenging conditions.

2.2 Track #2: Robust Sensor Placement

Theme & Motivation. This track challenges participants to design perception models that can adapt to diverse sensor placements in autonomous systems. Autonomous vehicles rely heavily on the accurate placement of sensors such as LiDARs and cameras for robust perception. However, the optimal positioning of these sensors can vary based on factors like vehicle design, operational environments, and the need to withstand adverse conditions such as sensor malfunctions, weather disruptions, and occlusions. Poor sensor placement can significantly degrade a vehicle's ability to perceive its surroundings, making it difficult to detect objects, estimate depth, or plan safe trajectories.

Participants will be tasked with developing algorithms that can adapt to and optimize sensor placements, ensuring high-quality perception across a wide range of environmental conditions. This track not only emphasizes sensor resilience but also focuses on generalizing perception models to multiple configurations and sensor layouts. By improving the adaptability and robustness of perception systems, this track aims to advance the development of safer, more reliable autonomous vehicles capable of operating in challenging real-world scenarios.

Datasets. The primary dataset used in this track is *CARLA-nuScenes* [12] from the Place3D benchmark, which combines simulated data from the CARLA driving simulator with nuScenes-style annotations and sensor setups. This dataset includes multiple sensor configurations (*e.g.*, multi-LiDAR and multi-camera setups) and is annotated with a wide range of environmental conditions and adverse scenarios such as sensor noise, occlusions, and varying lighting conditions. It also includes corresponding sensor placement information to test model adaptability across different layouts.

Baseline Model. The baseline models for this track are:

- BEVFusion-L [13]: A bird's eye view fusion model that integrates camera and LiDAR data to achieve 3D object detection. This model will serve as the baseline for evaluating sensor placement's impact on 3D object detection tasks.
- MinkUNet [4]: A voxel-based convolutional neural network used for LiDAR semantic segmentation. MinkUNet's performance under different sensor configurations will provide a baseline for LiDAR perception tasks in this track.

These baseline models provide a strong foundation for participants to optimize sensor configurations and improve perception quality, even under non-ideal or degraded sensor placements.

Evaluation Metrics. The performance of models in this track will be assessed using metrics that measure the quality of 3D perception across different sensor placements. The following metrics are used for evaluating the robustness and adaptability of the models:

- Mean Average Precision (mAP): Evaluates the accuracy of the object detection model in identifying and localizing objects within the scene.
- Mean Intersection over Union (mIoU): Measures the performance of the semantic segmentation model by calculating the overlap between the predicted and ground truth segmentation masks.
- Variance of Perception Performance across Configurations: Evaluates the consistency and robustness of the models when exposed to different sensor placements. Models that can adapt to various sensor setups without significant performance degradation will score higher.
- Variance of Perception Performance under Adverse Conditions: Tests the models' ability to maintain perception quality under challenging conditions such as sensor failures, occlusions, or environmental disturbances.

The evaluation will focus on the models' ability to generalize perception performance across different sensor placements and their resilience to adverse conditions, ensuring that the system can maintain robust perception capabilities in diverse operational scenarios.

2.3 Track #3: Robust Dense SLAM

Theme & Motivation. This track challenges participants to develop dense simultaneous localization and mapping (SLAM) models that maintain high accuracy under noisy sensor inputs, such as corrupted RGB-D video data. In real-world applications, SLAM systems must operate in dynamic environments where they are exposed to various challenges, including sensor noise, lighting variations, rapid camera movements, and dynamic objects. These conditions can introduce significant errors in trajectory estimation and 3D reconstruction quality, which in turn can compromise robot navigation and decision-making.

Participants in this track are tasked with building robust SLAM systems capable of generating high-quality 3D reconstructions and accurate camera trajectories, despite the presence of noise in the RGB-D inputs. This track focuses on advancing SLAM technology for autonomous systems, aiming to create models that can be deployed in harsh and unpredictable environments. Robust SLAM is essential for a wide range of robotics applications, including drone navigation, indoor robotics, and augmented reality, where accurate localization and 3D reconstruction are critical for system reliability.

Datasets. The dataset used in this track is *Robust-Ego3D* [20], a challenging RGB-D dataset designed specifically for benchmarking SLAM algorithms under noisy conditions. It contains RGB-D video streams with artificially introduced perturbations such as sensor noise, dynamic lighting, rapid motion, and environmental challenges (*e.g.*, occlusions and dynamic objects). It provides ground truth for both camera trajectories and 3D reconstructions, allowing for a comprehensive evaluation of the models' robustness in handling corrupted inputs.

Baseline Model. The baseline model for this track is CorrGS [20], a state-of-the-art dense neural RGB-D SLAM system that leverages geometric and photometric information for accurate 3D reconstruction and trajectory estimation.

Evaluation Metrics. The performance of models in this track will be evaluated based on their ability to accurately estimate camera trajectories and reconstruct 3D environments under noisy inputs. The following metrics are used:

- Absolute Trajectory Error (ATE): Measures the deviation between the predicted and ground truth camera trajectories. A lower ATE indicates more accurate localization.
- Peak Signal-to-Noise Ratio (PSNR): Evaluates the quality of the RGB reconstruction, with higher PSNR values indicating better reconstruction fidelity.
- L1 Loss in Depth: Measures the pixel-wise error in depth reconstruction, providing a quantitative assessment of the accuracy of the predicted depth maps.
- Frames Per Second (FPS): While not directly contributing to the primary ranking, models that achieve high FPS without sacrificing accuracy will be considered for the Innovative Award. This metric provides insights into the efficiency of the models when deployed in real-time applications.

The main evaluation criterion will be the Absolute Trajectory Error (ATE), which directly measures the accuracy of the SLAM model's localization. In cases of tied ATE, secondary metrics such as PSNR and depth L1 loss will be used to further rank the models. This ensures a comprehensive evaluation of both the localization and reconstruction capabilities of the SLAM models under noisy, real-world conditions.

2.4 Track #4: Robust Cross-View Matching

Theme & Motivation. This track focuses on developing models capable of robust cross-view matching. These models are specifically designed for scenarios where input data is captured from drastically different viewpoints, such as aerial (drone or satellite) and ground-level images. Cross-view matching has broad applications in geolocalization, autonomous navigation, and remote sensing, where it is essential for aligning data from different perspectives to the same geographic location. Natural environmental changes, lighting variations, occlusions, and the vast scale differences between images captured from different viewpoints further complicate the challenge.

Participants are tasked with creating models that can effectively match corresponding visual and textual elements across these differing views, even under the presence of corruption such as blurriness, occlusion, or environmental noise. The track emphasizes the importance of matching accuracy, robustness to input noise, and adaptability to natural challenges such as seasonal changes, weather, or lighting conditions. This work is critical for applications such as drone-based navigation, AR-guided geolocalization, and robust real-time geospatial recognition systems, where reliability under varying conditions is paramount.

Datasets. The primary dataset used in this track is *GeoText-1652* [5], an extension of the *University-1652* [21] dataset with fine-grained textual annotations added for cross-view matching tasks. This dataset includes a diverse range of images and associated textual descriptions from aerial (drone and satellite) and ground-level viewpoints. It covers 33 universities in the training set and 39 universities in the test set, ensuring no overlap in test and training classes. For each location, the dataset includes corresponding images and bounding box annotations from drone, satellite, and ground-level views, along with textual descriptions of both the global scene and specific object pairs.

Baseline Model. The baseline model for this track is a custom model built for cross-view matching. This baseline integrates visual and textual representations using a multi-modal approach, capable of aligning both image-to-image and text-to-image pairs across different viewpoints. The model leverages transformer-based architectures for matching tasks and employs contrastive learning to align visual and textual features across different views.

Evaluation Metrics. The performance of the models in this track will be evaluated based on their ability to match images and text across viewpoints, ensuring accurate retrieval and robust handling of input variations. The following metrics are used:

- Text Query (Text-to-Image Retrieval): Measures the accuracy of retrieving the correct image given a text description. Higher Recall@K indicates a more effective match between text descriptions and visual content.
- Image Query (Image-to-Text Retrieval): Measures the accuracy of retrieving the correct text description given an image. This evaluates the model's ability to generate meaningful textual matches for aerial or ground-view images.
- Bounding Box Prediction Accuracy: Evaluates how well the model predicts bounding boxes for specific objects across views, crucial for tasks like object localization and geolocalization.
- Spatial Relation Classification Accuracy: Assesses the model's ability to understand spatial relationships between objects, such as “the building is next to the courtyard” or “the tower is behind the library”, providing a deeper level of scene comprehension.

These metrics collectively evaluate the models' ability to robustly match and retrieve across different views, despite changes in scale, perspective, and environmental noise. Robustness to natural corruption is particularly critical, as cross-view matching systems must handle significant variations in input quality while still producing accurate matches.

2.5 Track #5: Robust Sensor Adaptation

Theme & Motivation. This track focuses on the development of robust perception models that can seamlessly adapt across different robotic platforms, including vehicles, drones, and quadrupeds. The primary challenge lies in ensuring that perception models, trained on one platform (in this case, vehicles), can generalize effectively to other platforms (such as drones and quadrupeds) without the need for additional ground truth labels or retraining. The motivation for this track stems from the increasing use of heterogeneous robotic systems in applications like autonomous driving, aerial surveillance, and legged navigation in challenging environments. Each of these platforms presents distinct challenges in terms of sensor placement, environmental interaction, and data quality, making robust adaptation critical for real-world applications.

Participants will be tasked with developing models that can effectively adapt perception tasks, specifically semantic segmentation, across platforms that use different sensor configurations and movement dynamics. These models must be able to generalize from vehicle-based sensor data (RGB, LiDAR, event cameras, IMU/GPS) to the more complex scenarios presented by drones and quadrupeds, which have different fields of view, motion patterns, and sensor placements. This track addresses the critical need for sensor adaptation across platforms, ensuring that models are not limited by the specific characteristics of the platform they were trained on, and can perform effectively in diverse robotic ecosystems.

Datasets. The dataset used in this track is *M3ED* [2], which provides a comprehensive collection of multi-sensor data across different platforms. This dataset includes event camera data, RGB images, LiDAR point clouds, and IMU/GPS data collected across three platforms: vehicles, drones, and quadrupeds. The dataset covers a wide range of environments, including urban areas, forests, and industrial zones, with varying weather conditions and lighting. The diversity of this dataset makes it ideal for studying sensor adaptation, as it includes data captured from platforms with different motion dynamics and sensor placements.

Baseline Model. The baseline model for this track is ESS [17], which focuses on event-based semantic segmentation. The baseline is designed to handle event camera data for semantic segmentation, leveraging temporal information from event streams to improve performance under various lighting and motion conditions. The model uses a combination of convolutional neural networks and spatiotemporal transformers to process event data and produce high-resolution semantic segmentation maps. The model is trained exclusively on vehicle data, using ground truth labels from *M3ED* for supervised learning. The primary challenge is adapting this model to perform well on drone and quadruped data, where the input distribution differs due to varying sensor placements and motion patterns, without the benefit of ground truth during testing.

Evaluation Metrics. The performance of the models in this track will be assessed based on their ability to adapt across platforms, with a specific focus on semantic segmentation. The primary evaluation metrics are:

- Mean Intersection over Union (mIoU): This is the core metric for evaluating the quality of semantic segmentation. The mIoU score will be computed based on the overlap between the predicted segmentation maps and ground truth (where available during training). For the final evaluation, mIoU will be calculated for the models' performance on the test platforms (drone and quadruped) without using ground truth labels during testing.

In addition to the mIoU score, the track will also consider:

- Cross-Platform Generalization: The models' ability to generalize from vehicle-based training to drone and quadruped testing will be a key factor. This will be evaluated based on how well the models can maintain high segmentation performances when deployed on platforms with different sensor setups and environmental interactions.

The objective is to ensure that models can robustly adapt their perception capabilities across different robotic platforms, enabling broader applicability in real-world multi-platform environments.

3 Time Schedule

The **RoboSense** Challenge will follow a structured timeline to ensure participants have sufficient time to develop and refine their models for robust robot perception in challenging conditions. The schedule is outlined as follows:

- **Early Preparation (January 2025 – March 2025)**
  During this period, the competition toolkit, datasets, and instructions will be made available on the official competition website. Participants will be able to register their teams and start preparing for the challenge.
  - March 2025: Announcement of the competition and official website launch.
  - March 1st, 2025 – March 14th, 2025: Release of training, validation, and evaluation datasets, including associated documentation, benchmark baselines, and competition toolkit.

  - March 15th, 2025: Access to the training and evaluation data will be provided through the competition toolkit and official GitHub repository. Competitors can familiarize themselves with the data and start model development.

- **Phase One: Competition Start (March 2025 – June 2025)**
  The competition will officially begin on March 15th, 2025, and participants will have access to both the training and evaluation datasets. The CodaLab competition server will go live, enabling participants to test their models and submit their results for the first phase of the competition.

  - March 15th, 2025: CodaLab competition server goes live for model evaluation. Competitors can submit their results for continuous feedback and performance evaluation on the benchmark.
  - June 30th, 2025: Phase One submission deadline. Shortlisted teams from this phase will be invited to participate in the final round (Phase Two).

- **Phase Two: Final Submission (July 2025 – September 2025)**
  Teams shortlisted from Phase One will move on to Phase Two, where they will fine-tune their models and submit final results, including code and documentation. The final phase will focus on evaluating the robustness of the models under challenging conditions such as sensor corruption and environmental noise.

  - September 30th, 2025: Deadline for Phase Two submissions. Participants are required to submit their final models, along with the corresponding code and documentation, by this date.

- **Award Decision (October 19th, 2025)**
  Following the submission deadline, the evaluation committee will assess the final results and determine the top-performing teams. The award decision will be announced during IROS 2025.

  - October 19th, 2025: Official announcement of the award-winning teams, held in conjunction with IROS 2025 in Hangzhou, China. A formal presentation will showcase the best-performing models and innovative solutions. A competition report summarizing the solutions will be made available after the event, highlighting the key achievements and insights gained from the competition.

This schedule provides a comprehensive framework for organizing the competition, ensuring that teams have adequate time for development, evaluation, and final submissions leading up to the IROS 2025 conference.

4 Potential Participant Teams

The RoboSense Challenge invites participation from a broad range of researchers, engineers, and students from both academia and industry, with interests in developing resilient robot perception systems. The competition's emphasis on robust perception in challenging environments, including sensor adaptation, cross-view matching, SLAM, and multi-platform integration, makes it particularly relevant to the fields of robotics, computer vision, and autonomous systems.

We encourage participation from the following groups:

- **Academia:** Researchers, graduate students, and teaching staff engaged in relevant research areas such as robot perception, vision-language models, SLAM, cross-view matching, and sensor adaptation. We especially welcome individuals or teams working on projects or dissertations related to robust perception in challenging conditions.

- **Industry Professionals:** Engineers, developers, and researchers from leading companies in the fields of robotics, automation, autonomous driving, drone technology, and sensor systems. The competition's focus on real-world robustness aligns with the industrial need for reliable perception systems across diverse platforms and conditions.

- **Autonomous Driving & Drone Development Teams:** Teams from the autonomous driving and drone navigation sectors who are actively developing perception systems that can handle adverse conditions like sensor failures, weather perturbations, or challenging sensor placements.

- **Students & Educators:** Undergraduate and graduate students participating in relevant course projects or robotics competitions. Educators interested in integrating real-world perception challenges into their curriculum are also encouraged to participate. The competition provides an opportunity to engage with cutting-edge technologies while contributing to the development of practical, robust perception solutions.

- **Multi-Disciplinary Teams:** Given the competition's multi-modal approach, which spans areas such as SLAM, cross-view matching, and sensor adaptation across platforms (vehicles, drones, quadrupeds), interdisciplinary teams with expertise in sensor fusion, machine learning, and hardware-software co-design are highly encouraged to join.

Based on previous participation in similar competitions, including this year's RoboDrive Challenge, we anticipate strong interest from both academic and industrial teams. **We expect around 50 to 100 teams to participate across the five tracks** of the RoboSense Challenge. We believe this competition will offer a unique platform for collaboration and innovation, addressing key challenges in robust robot perception under real-world conditions, while fostering connections between academia and industry.

5 Awards Planning & Estimated Budget

To recognize and encourage top-performing teams in the RoboSense Challenge, we plan to provide a series of awards across all five tracks. These awards will be based on the performance of the submissions evaluated using the competition's robust perception metrics, specific to each track's objectives. The awards aim to acknowledge excellence in various dimensions, including technical innovation, robustness, and generalization capabilities.

5.1 Awards Breakdown

The following awards are planned for each track of the competition:

- **First Prize (Five Awardees)**
  Cash award of USD $1,000 per track, for a total of USD $5,000 across all five tracks. The First Prize will be awarded to the top-performing team in each track based on the evaluation metrics, including robustness under sensor failures, environmental noise, and other challenging conditions.

- **Second Prize (Five awardees)**
  Cash award of USD $600 per track, for a total of USD $3,000 across all five tracks. The Second Prize will be awarded to teams whose submissions demonstrate excellent performance but rank just below the first prize winners.

- **Third Prize (Five Awardees)**
  Cash award of USD $400 per track, for a total of USD $2,000 across all five tracks. The Third Prize will be awarded to teams that show strong performance and consistency across various challenges but are ranked below the second prize winners.

- **Innovative Award**
  Certificates will be given to up to two teams per track for innovative approaches or novel solutions. The Innovative Award recognizes creativity and out-of-the-box thinking that may not necessarily rank highest in performance metrics but demonstrate promising directions for future research.

All winners will also receive certificates acknowledging their achievements, and the competition results will be presented during the IROS 2025 event. The winners' solutions will be included in the competition's official report, providing further visibility within the academic and industrial communities.

5.2 Estimated Budget

Given the award structure outlined above, the total budget for awards is estimated as follows:

- First Prize: $\$1,000 \times 5 = \$5,000$
- Second Prize: $\$600 \times 5 = \$3,000$
- Third Prize: $\$400 \times 5 = \$2,000$
- Innovative Award: Certificates (no cash award)

Total Award Budget: USD $\$\mathbf{10,000}$

5.3 Sponsorship & Funding

We are delighted that the RoboSense Challenge has secured a sponsorship from HUAWEI UK Research Center, which has generously provided USD $\$\mathbf{10,000}$ to support this competition. This sponsorship will be allocated toward competition prizes, further enhancing the incentives for participants to develop innovative and robust solutions in robotic perception.

While the award structure and budget are now partially finalized, we are actively seeking additional sponsorships to further enhance the competition and provide even greater support for participants. The awards and recognition offered to participants present an excellent opportunity for sponsors to engage with a global community of top researchers and engineers at the forefront of robotics and computer vision.

We are also maintaining communication with previous sponsors from our last two competitions, and we are optimistic about securing additional funding for this year's event.

In addition to cash prizes, all sponsors will be acknowledged in competition-related publications, the official competition website, and the award ceremony at IROS 2025. This visibility ensures sponsors receive significant recognition within the robotics and AI communities while aligning with the mission of advancing robust robotic perception systems.

6 Special Requests

We aim to host the RoboSense Challenge in person at IROS 2025, Hangzhou, China, allowing for direct interactions, presentations, and demonstrations from participating teams. However, to ensure flexibility in response to any unforeseen circumstances, such as travel restrictions or other limitations, we have prepared both an **on-site** and an **online** plan in a hybrid manner.

6.1 On-Site Plan

For the in-person component, we kindly request the following arrangements to facilitate the smooth running of the competition and award ceremony:

- Spaces: A 5m $\times$ 10m rectangular area for hosting the workshop. A dedicated workshop room is preferred.

- Projector: For presentation and results announcement.
- WiFi Access: Stable internet access for teams and presenters.
- Power Outlets: Sufficient power supply for teams' laptops and equipment.
- Microphones: At least one microphone for presentations and discussions.
- Desks and Chairs: Seating arrangements for participants and audience members.
- Poster Boards: 5–10 poster stands and supporting frames for teams to display their work (optional).
- Volunteers: 1–3 volunteers to assist with registration, setup, and event logistics (optional).

6.2 Online Plan

- Zoom Link: An official Zoom link to host the award ceremony and presentations for the winning teams.
- Virtual Environment: Utilize platforms like *Gather.Town* to facilitate team networking and virtual poster sessions.
- Certificates: Provide digital certificates for the winning teams, which can be shared and acknowledged online.

7 History of the Competition

The **RoboSense** Challenge at IROS 2025 builds upon a rich history of organizing successful academic competitions, particularly the RoboDrive and RoboDepth challenge series. These competitions have played a pivotal role in advancing the field of robust robot perception, with a specific focus on out-of-distribution (OoD) scenarios, sensor failures, and challenging real-world environments.

7.1 The RoboDepth Challenge @ ICRA 2023

RoboDepth [7] was first introduced at ICRA 2023, London, UK, and has since become a popular benchmark for testing and improving the OoD robustness of depth estimation models. It was developed based on the KITTI-C and NYUDepth2-C benchmarks [8]. This competition was the first of its kind, targeting the robustness of monocular depth estimation under various environmental challenges such as weather, lighting changes, and sensor corruption.

Key Achievements of RoboDepth

- Tracks: Two stand-alone tracks focused on self-supervised and fully-supervised depth estimation.
- Participation: Over 200 participants with nine top-performing solutions.
- Technological Contributions: Key innovations included adversarial training, diffusion-based noise suppression, hierarchical feature enhancement, and multi-modal sensor fusion.
- Public Resources:
  - Website: https://robodepth.github.io.
  - Competition report: https://arxiv.org/abs/2307.15061.
  - Video recordings: https://www.youtube.com/watch?v=mYhdTGiIGCY&list=PLxxrIfcH-qBGZ6x_e1AT2_YnAxiHIKtkB.

7.2 The RoboDrive Challenge @ ICRA 2024

Following the success of RoboDepth, we launched the RoboDrive Challenge [9] at ICRA 2024 in Yokohama, Japan, with a strong focus on enhancing the robustness of perception systems for autonomous driving. This competition specifically targeted four key tasks critical for reliable navigation: bird's eye view detection, HD map segmentation, semantic occupancy prediction, and multi-view depth estimation. Each of these tasks is essential for ensuring accurate decision-making in dynamic, real-world driving environments.

Key Achievements of RoboDrive

- Tracks: Five distinct tracks focusing on bird's eye view detection, map segmentation, semantic occupancy prediction, and multi-view depth estimation.
- Participation: 140 teams from 93 institutes, with nearly 1,000 submissions. A total of 15 top-performing solutions were selected.
- Technological Contributions: Participants introduced innovations in multi-sensor fusion, data augmentation, and sensor error correction, significantly improving perception robustness.
- Public Resources:
  - Website: https://robodrive-24.github.io.
  - Competition report: https://arxiv.org/abs/2405.08816.
  - Video recordings: https://youtu.be/83_FxdXNNQk.

7.3 Expanding the Scope in 2025

The **RoboSense** Challenge expands on the success of RoboDrive and RoboDepth by introducing five new tracks that focus on various aspects of **robust robot sensing across different platforms and environments**. These tracks aim to address critical challenges in sensor placement, SLAM, cross-view matching, and sensor adaptation, offering participants a broad range of tasks designed to push the boundaries of resilient robot sensing systems.

The lessons learned and technologies developed through the RoboDrive and RoboDepth challenges provide a strong foundation for the **RoboSense** Challenge. We anticipate this new competition will further elevate research and innovation in robust robot perception, providing new insights and advancing the state of the art in resilient autonomous systems.

8 Organizers

The organizing team of the **RoboSense** Challenge comprises experts with extensive experience in the relevant topics, representing diverse regions including North America, East Asia, Europe, and Southeast Asia.

- **Lingdong Kong** (Contact: `lingdong@comp.nus.edu.sg`)
  Lingdong is a Ph.D. candidate in the School of Computing, Department of Computer Science, at the National University of Singapore. His research interests include 3D scene perception, generation, and understanding. Personal website: `https://ldkong.com`.
  Experience. Lingdong was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024, the *RoboDepth Challenge* [link] at ICRA 2023, and the *4th SenseHuman Workshop* [link] and the *PointCloud-C Challenge* [link] at ECCV 2022.

- **Shaoyuan Xie** (Contact: `shaoyux@uci.edu`)
  Shaoyuan is a Ph.D. student in the Donald Bren School of Information and Computer Sciences at the University of California, Irvine. He worked in the Computational Cognition, Vision, and Learning Labs at Johns Hopkins University as a research intern. His research interests include computer vision and trustworthy machine learning systems. Personal website: `https://daniel-xsy.github.io`.
  Experience. Shaoyuan was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024 and the *RoboDepth Challenge* [link] at ICRA 2023.

- **Ye Li** (Contact: `yeyli@umich.edu`)
  Ye is a Master's student in the Robotics Department at the University of Michigan, Ann Arbor. His research interests include 3D scene understanding and optimization theory. Personal website: `https://ywyeli.github.io`.

- **Xiaohao Xu** (Contact: xiaohaox@umich.edu)
  Xiaohao is a Ph.D. candidate in the Robotics Department at the University of Michigan, Ann Arbor. He worked in Microsoft Research Asia as a research intern. His research interests include robust perception, anomaly detection, and video understanding. Personal website: https://robotics.umich.edu/profile/xiaohao-xu.

- **Feng Xue** (Contact: fengxe@umich.edu)
  Feng Xue is a Master's student in the Robotics Department at the University of Michigan, Ann Arbor. He worked at ROMALab at the Southern University of Science and Technology as a research assistant. His research interests include SLAM and vision-based perception.

- **Meng Chu** (Contact: chumeng@pjlab.org.cn)
  Meng is a research assistant at Shanghai AI Laboratory. He obtained his Master's degree from the National University of Singapore (NUS). He worked in NeXT++, a joint research center between NUS, Tsinghua University, and Southampton University, for two years. His research interests include vision and language, representation learning, and embodied intelligence.

- **Yaru Niu** (Contact: yarun@andrew.cmu.edu)
  Yaru is a Ph.D. candidate in the Safe AI Lab at Carnegie Mellon University. He received his M.S. degree from the Georgia Institute of Technology and B.Eng. from the South China University of Technology. His research interest lies at the intersection of lies at the intersection of robotics, machine learning, and multi-agent systems. His research goal is to build methodologies for collaborative, interpretable, and reliable intelligent robotic systems that can interact with complex environments around them. Personal website: https://yaruniu.com/.
  Experience. Yaru was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024 and the *RoboDepth Challenge* [link] at ICRA 2023.

- **Hanjiang Hu** (Contact: hanjianh@cs.cmu.edu)
  Hanjiang is a Ph.D. candidate in the Department of Electrical and Computer Engineering (ECE) and a Master of Machine Learning (MSML) student at Carnegie Mellon University. He received his B.S. and M.S. degrees from Shanghai Jiao Tong University. His research focuses on safety and robustness at the intersection of robotics, control theory, and machine learning. He is also interested in the adversarial and OOD robustness caused by semantic transformation in real-world robotics and autonomous driving applications. Personal website: https://cs.cmu.edu/~hanjianh.
  Experience. Hanjiang was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024, the *RoboDepth Challenge* [link] at ICRA 2023, the *SeasonDepth Prediction Challenge* [link] at ICRA 2022 and IROS 2022, and the *Trustworthy Autonomy and Robotics Workshop* [link] at ICRA 2022.

- **Zhedong Zheng** (Contact: zhedongzheng@um.edu.mo)
  Zhedong is an Assistant Professor at the University of Macau. He received the Ph.D. degree from the University of Technology Sydney in 2021 and the B.S. degree from Fudan University in 2016. He was a postdoctoral research fellow at the School of Computing, National University of Singapore. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He served as the senior PC for IJCAI and AAAI, and the area chair for ACM MM'24.
  Experience. Zhedong was within the organizing team of the *UAVM Workshop* [link] and the *MMGR Workshop* [link] at ACM MM 2023 and ACM MM 2024.

- **Xiaonan (Sean) Huang** (Contact: xiaonanh@umich.edu)
  Xiaonan is an Assistant Professor in the Robotics Department at the University of Michigan, Ann Arbor. He leads the Hybrid Dynamic Robotics Lab (HDRL), where his research focuses on the design, control, and SLAM for soft robots operating in unstructured environments. He earned his Bachelor's degree from Harbin Institute of Technology and his Master's and PhD degrees from Carnegie Mellon University. Dr. Huang has served as an Associate Editor for prominent robotics conferences, including IROS, BioRob, and RoboSoft.

- **Benoit R. Cottereau** (Contact: benoit.cottereau@cnrs.fr)
  Benoit is a CNRS Research Director in computational neurosciences and vision affiliated to the Brain and Cognition Research Center (Cerco, CNRS UMR 5549, Toulouse) in France and to the International Research Laboratory on Artificial Intelligence (IPAL, IRL 2955) in Singapore where he co-leads a research axis on efficient AI. He is also an Adjunct Associate Professor in the School of Computing, Department of Computer Science, at the National University of Singapore (NUS). His research interests include visual processing in artificial systems with a focus on visual scene understanding, e.g., semantic segmentation, depth estimation, and motion prediction. Personal website: https://ipal.cnrs.fr/benoit-cottereau-personal-page.
  Experience. Benoit was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024 and the *RoboDepth Challenge* [link] at ICRA 2023.

- **Lai Xing Ng** (Contact: ng_lai_xing@i2r.a-star.edu.sg)
  Lai Xing is a Scientist from the Visual Intelligence Department at the Institute for Infocomm Research, A*STAR, Singapore. His research interests include computer vision, visual perception, visual inspection, cognitive task reasoning, scene understanding, and the development of AR/VR applications. Personal website: https://ipal.cnrs.fr/lai-xing-ng.
  Experience. Lai Xing was within the organizing team of the *RoboDrive*

Challenge [link] at ICRA 2024 and the *RoboDepth Challenge* [link] at ICRA 2023.

- **Wei Tsang Ooi** (Contact: ooiwt@comp.nus.edu.sg)
Wei Tsang is currently an Associate Professor of Computer Science in the School of Computing at the National University of Singapore (NUS). He is currently a co-director of IPAL, a Franco-Singaporean International Research Lab. His research focuses on building interactive multimedia systems that provide a high-quality experience, are robust, and are resource efficient. Personal website: https://www.comp.nus.edu.sg/cs/people/ooiwt.
Experience. Wei Tsang is an active member of the ACM SIGMM community and has served on the organizing committee of many conferences and workshops, including service as the TPC co-chair of ACM MMSys 2014 and ACM MM 2019; General Chair of NOSSDAV 2009, ACM MMSys 2015, and MMVE 2018. Wei Tsang is also part of the team that runs a series of *ACM Multimedia Grand Challenges* in video streaming in 2019, 2021, and 2022. Wei Tsang was part of the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024 and the *RoboDepth Challenge* [link] at ICRA 2023.

- **Wenwei Zhang** (Contact: zhangwenwei@pjlab.org.cn)
Wenwei received the Ph.D. degree from Nanyang Technological University, Singapore, in 2024. He is a Research Scientist at Shanghai AI Laboratory. His research interests include 2D & 3D object detection and segmentation. He is a core maintainer of MMDetection, MMDetection3D, and MMCV in OpenMMLab, an open-source computer vision platform. Personal website: https://zhangwenwei.cn.
Experience. Wenwei was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024.

- **Liang Pan** (Contact: panliang1@pjlab.org.cn)
Liang received the Ph.D. degree from the National University of Singapore, Singapore, in 2019. He is a Research Scientist at Shanghai AI Laboratory. Previously, he was a research fellow at S-Lab, Nanyang Technological University, Singapore, from 2020 to 2023, and a research fellow at the Advanced Robotics Center, National University of Singapore, in 2019. His research interests include shape analysis, deep learning, and 3D humans.
Experience. Liang was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024.

- **Ziwei Liu** (Contact: ziwei.liu@ntu.edu.sg)
is an Assistant Professor in the College of Computing and Data Science at Nanyang Technological University, Singapore. Previously, he was a senior research fellow at the Chinese University of Hong Kong and a postdoctoral researcher at the University of California, Berkeley. He received the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong,

China, in 2017. His research revolves around computer vision, computer graphics, machine learning, and robotics. He was the recipient of the MIT Technology Review Innovators under 35 Asia Pacific, ICBS Frontiers of Science Award, Microsoft Young Fellowship, Hong Kong Ph.D. Fellowship, ICCV Young Researcher Award, WAIC Yunfan Award, HKSTP Best Paper Award, and CVPR Best Paper Award Candidate. Personal website: `https://liuziwei7.github.io`.

Experience. Ziwei was within the organizing team of the *RoboDrive Challenge* [link] at ICRA 2024, the *OmniObject3D Challenge* [link] at ICCV 2023, and the *VBench Challenge* [link] at CVPR 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 4

[2] 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4016–4023, 2023. 10

[3] Mu Chen, Zhedong Zheng, and Yi Yang. Transferring to real-world layouts: A depth-aware framework for scene adaptation. *ACM MM*, 2024. 1

[4] 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*, pages 3075–3084, 2019. 6

[5] Meng Chu, Zhedong Zheng, Wei Ji, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatially relation matching. *European Conference on Computer Vision*, 2024. 9

[6] 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. 1

[7] 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. 16

[8] 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*, 2023. 16

[9] 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. 1, 17

[10] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1738–1764, 2022. 1

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, pages 19730–19742, 2023. 5

[12] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your lidar placement optimized for 3d scene understanding? *Advances in Neural Information Processing Systems*, 37, 2024. 6

[13] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, pages 2774–2781, 2023. 6

[14] Armin Masoumian, Hatem A. Rashwan, Julián Cristiano, M. Salman Asif, and Domenec Puig. Monocular depth estimation using deep learning: A review. *Sensors*, 22(14):5353, 2022. 1

[15] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5314–5334, 2022. 1

[16] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *European Conference on Computer Vision*, 2024. 5

[17] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. *European Conference on Computer Vision*, pages 341–357, 2022. 10

[18] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *IEEE International Conference on 3D Vision*, pages 11–20, 2017. 1

[19] 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. *arXiv preprint arXiv:2405.17426*, 2024. 4

[20] Xiaohao Xu, Tianyi Zhang, Sibo Wang, Xiang Li, Yongqi Chen, Ye Li, Bhiksha Raj, Matthew Johnson-Roberson, and Xiaonan Huang. From perfect to noisy world simulation: Customizable embodied multimodal perturbations for slam robustness benchmarking. *arXiv preprint arXiv:2406.16850*, 2024. 7, 8

[21] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM international conference on Multimedia*, pages 1395–1403, 2020. 9