aiMotive Dataset: A Multimodal Dataset for Robust Autonomous Driving with Long-Range Perception

Tamás Matuszka, Iván Barton, Ádám Butykai, Péter Hajas, Dávid Kiss
Domonkos Kovács, Sándor Kunsági-Máté, Péter Lengyel, Gábor Németh
Levente Pető, Dezső Ribli, Dávid Szeghy, Szabolcs Vajna, Bálint Varga
aiMotive
Budapest, Hungary
https://aimotive.com

Abstract

Autonomous driving is a popular research area within the computer vision research community. Since autonomous vehicles are highly safety-critical, ensuring robustness is essential for real-world deployment. While several public multimodal datasets are accessible, they mainly comprise two sensor modalities (camera, LiDAR) which are not well suited for adverse weather. In addition, they lack far-range annotations, making it harder to train neural networks that are the base of a highway assistant function of an autonomous vehicle. Therefore, we introduce a multimodal dataset for robust autonomous driving with long-range perception. The dataset consists of 176 scenes with synchronized and calibrated LiDAR, camera, and radar sensors covering a 360-degree field of view. The collected data was captured in highway, urban, and suburban areas during daytime, night, and rain and is annotated with 3D bounding boxes with consistent identifiers across frames. Furthermore, we trained unimodal and multimodal baseline models for 3D object detection. Data are available at https://github.com/aimotive/aimotive_dataset.

1 Introduction

A large number of datasets for 3D object detection applied in autonomous driving (AD) have been released in the last few years (Geiger et al., 2012b; Chang et al., 2019; Huang et al., 2018; Pham et al., 2020; Patil et al., 2019; Caesar et al., 2020). The majority of datasets have the common property of including sensor data from different modalities, including cameras and LiDAR. In this way, a 360-degree field-of-view (FOV) can be covered around the ego vehicle. 3D object detection datasets can be split into different groups along the dimensions of the coverage around the ego vehicle and sensor redundancy. While numerous datasets are publicly available, they either do not provide sensor redundancy (i.e. coverage by at least two sensor modalities) which is essential for robust autonomous driving or rely only on camera and LiDAR sensors that are not perfectly applicable in adverse weather (see Table 1 for the properties of several popular datasets grouped based on sensor coverage and redundancy). This issue could be solved by utilizing radars which is a cost-effective sensor and is not affected by adverse environmental conditions (e.g. rain or fog). Furthermore, the annotation range does not exceed 80 meters (with a few exceptions) which is not sufficient for training long-range perception systems. The limitation of the annotation range can be explained by the fact that autonomous driving datasets mostly focus on urban environments while ensuring the ability to detect objects in distant regions is critical for highway assistants and therefore for autonomous driving.

In order to overcome the above-mentioned limitations, we release a multimodal dataset for robust autonomous driving with long-range perception. The collected dataset includes 176 scenes with synchronized and calibrated LiDAR, camera, and radar sensors covering a 360-degree field of view. The data was captured in diverse geographical areas (highway, urban, and suburban) and different time and weather conditions (daytime, night, rain). We provide 3D bounding boxes with consistent

Dataset	Ours	nuScenes	Lyft	Argo2	Waymo	ONCE	Radiate	RADIal	DENSE
Ann. Scenes	176	1000	366	1000	**1150**	581	61	91	N/A
Ann. LiDAR frames	26.5k	40k	46k	150k	**230k**	16k	44k	8.3	13.5k
3D boxes	427k	1.4M	1.3M	11.25M	**12M**	412k	203k	9.5k	100k
Classes	14	23	9	**30**	4	5	8	1	4
Camera	4	6	7	**9**	5	7	2	1	2
Radar	2	**5**	-	-	-	-	1	1	2
LiDAR	1	1	3	2	**5**	1	1	1	1
Range	**<200m**	<100m	<100m	**<200m**	<100m	<100m	<100m	<100m	<100m
Countries	3	2	1	1	1	1	1	1	**4**
Continents	**2**	**2**	1	1	1	1	1	1	1
Night	✓	✓	✗	✓	✓	✓	✓	✗	✓
Rain	✓	✓	✓	✓	✓	✓	✓	✗	✓
Highway	✓	✗	✗	✗	✗	✗	✓	✓	✓

Table 1: Comparison of relevant datasets. Middle group: datasets with redundant 360° sensor coverage, right group: datasets with 360° view without sensor redundancy. Range refers to the perception limit of the front and back region in the case of the middle group and the front area for the right group (ego vehicle is the origin).

identifiers across frames which enables the utilization of our dataset for 3D object detection and multiple object tracking tasks. The proposed dataset is published under CC BY-NC-SA 4.0 license, allowing the research community to use the gathered data for non-commercial research purposes. Our main contributions are the followings:

- We released a multimodal autonomous driving dataset with redundant sensor coverage including radars and 360° FOV.
- Our dataset has an extended annotation range compared to existing datasets allowing the development of long-range perception systems.
- We trained and benchmarked unimodal and multimodal baseline models.

By releasing our dataset and models to the public, we seek to facilitate research in multimodal sensor fusion and robust long-range perception systems.

2 Related Work

One of the most influential datasets is KITTI by Geiger et al. (2012b), which generated interest in 3D object detection in autonomous driving. The KITTI dataset contains 22 scenes recorded in Karlsruhe, Germany. The sensor setup consists of front cameras and a roof-mounted LiDAR. The perception range of the released dataset is less than 100 meters, and no 360-degree FOV is provided. In addition, the footage was recorded only in the daytime.

Several popular 3D object detection datasets provide a 360° FOV with sensor redundancy. nuScenes (Caesar et al., 2020) is the most similar dataset to our work, including full sensor redundancy for the entire sensor setup. However, a 32-beam LiDAR with a relatively sparse point cloud and limited perception range was used during the recording process, resulting in a shorter perception limit than 100 meters (i.e. there are no annotated objects with a distance larger than 100 meters from the ego vehicle at the moment when the given frame was annotated). The sensor data has been recorded in urban environments (Boston, USA, Singapore) and lacks footage on highways. Waymo Open Dataset(Sun et al., 2020) is the first large-scale autonomous driving 3D object detection data collection with 360° FOV, including more than 1000 scenes and 12M annotated objects. The main shortcoming of this dataset is the limited perception range and sensor suite. The recently released Argoverse2 Sensor (Wilson et al., 2021) dataset utilized the experiences gained from hosting several challenges using the Argoverse (Chang et al., 2019) dataset. Argoverse2 has a similar scale as Waymo Open Dataset but with an extended annotation range. The disadvantage of the dataset compared with our solution is the lack of radar sensor usage and the diversity of recording locations (see Table 1). Both Lyft Level 5 perception dataset (Kesten et al., 2019) and ONCE (Mao et al., 2021) have recordings from

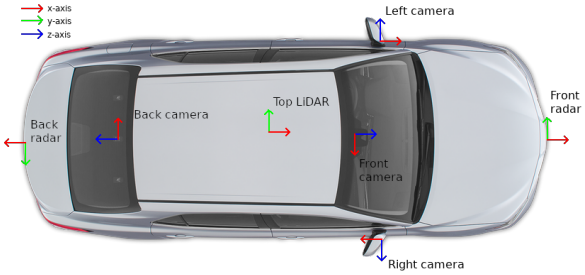


Figure 1: Sensor setup and coordinate systems.

Sensor	Details
1x LiDAR	Spinning, 64 beams, 10Hz capture frequency, 360° horizontal FOV, +15° to -25° vertical FOV, $\leq$ 200m range, ± 2 cm accuracy, up to 1.15 Million Points per Second
2x Radar	18 Hz capture frequency, 77GHz, -16° to +16° azimuth FOV, -7.5 to +7.5 elevation FOV, up to 175m distance, velocity accuracy of ±0.1 m/s
4x Camera	2x fisheye: 30 to 60 Hz capture frequency, 1/2.7" CMOS sensor of 1920 × 1080 resolution, 2x pinhole: 30 to 40 Hz capture frequency, 1/1.55" CMOS sensor of 2896 × 1876 resolution
1x GNSS+INS	up to 100 Hz measurement frequency, position accuracy of 100 mm (PPK), heading accuracy of 0.044°, roll&pitch accuracy of 0.009°

Table 2: Description of used sensors.

only one country, without utilizing any radars, and do not contain annotated objects in distant areas. Radiate (Sheeny et al., 2021) uses three different sensor modalities and contains a large number of annotated keyframes in adverse weather (e.g. fog, rain, snow). The paper's main contribution is the release of a high-resolution radar dataset. However, the perception range is limited (i.e. less than 100 m), and other sensor modalities are constrained (32-beam LiDAR with very sparse point cloud, only front camera with low-resolution images).

Another group of datasets also provides 360-degree coverage without sensor redundancy, which is essential for robust autonomous driving. RADIal(Rebut et al., 2022), similar to Radiate, employs a high-definition radar for sensing in 360°. The recorded data covers a wide range of geographical areas, however, the sensor setup is restricted to only three sensors. Furthermore, the dataset has a limited amount of annotated objects (less than 10k). DENSE (Bijelic et al., 2020) also focuses on data collected in severe weather. The paper describes a unique sensor setup including a thermal camera, gated cameras, and a spinning LiDAR. Even though a diverse set of sensors is mounted to the recording car, sensor redundancy is not ensured in the case of the dataset. Moreover, the annotated area is limited due to the challenging weather conditions.

As Table 1 summarizes, our dataset has an advantage over the existing related work. The proposed dataset combines full sensor redundancy with a long perception range in diverse environments, which is not provided by previously published 3D object detection datasets. Ensuring these properties are required for training neural networks which can serve as a base of robust autonomous driving software that can operate in different environments.

3 AIMOTIVE MULTIMODAL DATASET

Our multimodal dataset comprises 15s long scenes with synchronized and calibrated sensors. The dataset provides a 360° FOV using a redundant sensor layout where the area around the ego vehicle is recorded by at least two different sensors at the same time. Since the annotated 3D bounding boxes have consistent identifiers across frames, the dataset can be used for 3D object detection and multiple object tracking tasks. In addition, a considerable amount of annotations (about 25%) are located in the far-distance region ($\geq 75m$) concerning the ego vehicle. Due to this property and the redundant sensor setup, our dataset can facilitate research in multimodal sensor fusion and robust long-range perception systems.

3.1 DATA COLLECTION

The data was collected in three countries on two continents with four cars to provide a diverse dataset. The recordings have taken place in California, US; Austria; and Hungary using three Toyota Camry and one Toyota Prius. The recording phase of the footage was spread across a year to gather data in different seasons and weather conditions. As a result, our dataset consists of a diverse set of locations (highway, suburban, urban), times (daytime, night), and weather conditions (sun, cloud,

	Highway	Urban	Night	Rain
All	45.8%	35.4%	15.0%	3.8%
Train	46.3%	34.4%	16.6%	2.7%
Val	43.4%	39.4%	8.6%	8.6%

Table 3: Data distribution w.r.t weather and environment.

	Length	Width	Height
Car	4.20	1.93	1.63
Pedestrian	0.77	0.75	1.77
Motorcycle	2.02	0.89	1.53
Bicycle	1.58	0.82	1.54

Table 4: Average cuboid dimensions (m).

rain, glare). The data collection method has satisfied the requirements given by the Institutional Review Board approval.

3.2 Sensor Setup

Sensor layout. The data was recorded using a roof-mounted, rotating 64-beam LiDAR, four cameras, and two long-range radars, providing 360° coverage with sensor redundancy. The localization was based on a high-precision GNSS+INS sensor. Additional details can be found in Figure 1 and Table 2.

Synchronization. All of the recorded sensor data are synchronized. The LiDAR and radars share the same timestamp source. Our cameras capture images using the rolling shutter method, which scans the environment rapidly instead of capturing the image as a snapshot of the entire scene at a single time moment. Since the used cameras capture the scene row by row, the camera timestamp is approximately the exposure time when the middle row is captured.

Coordinate system. The dataset uses five coordinate systems, namely global, body, radar, camera, and image coordinate systems. We have used ECEF (Snay & Soler, 1999) as the global coordinate system and provided a 6-DOF ego-vehicle pose for each annotated frame. The reference coordinate system used for defining the annotated objects is called the body coordinate system that is attached to the vehicle body. The origin is the projected ground plane point under the center of the vehicle's rear axis at nominal vehicle body height and zero velocity. The radar coordinate system uses the same axes as the body coordinate system (x-axis positive forward, y-axis positive to the left, and z-axis positive upwards). The LiDAR point cloud was transformed into the body coordinate system as a preprocessing step. The origin of the camera coordinate system is the camera's viewpoint and the axes are defined the same as the OpenCV(Kaehler & Bradski, 2016) camera coordinate system (x-axis positive to the right, y-axis positive downwards, z-axis positive forward). Camera-to-body and radar-to-body transformations can be performed using camera and radar extrinsic matrices. We utilized OpenCV's image coordinate system for rendering annotations using intrinsic matrices to project from camera coordinates to image coordinates.

3.3 Ground Truth Generation

We used two approaches for generating ground truth labels: an automatic annotation method for training data generation and manual annotation for creating validation data. The automatic annotation relies on LiDAR measurements and searches possible candidates in the entire point cloud of a 15 s long sequence. Non-causal object tracking (i.e. both directions in time) including the association of new detections to the existing tracks is realized in the 2.5D descriptor space with the joint probability distribution of the modeled detection uncertainty and object dynamics. Utilizing the informative point cloud along with the physical constraints, the consecutive detections (i.e. positions and orientations) of the same object can be optimized recursively. In this way, the point cloud of a given object can be accumulated from different views. As the tracked object's trajectory becomes more accurate with the optimization steps, the model point cloud of the detected object becomes sharp, thus, a bounding box can be fitted on it. The annotated sequences were manually quality-checked based on multiple criteria. This inspection checks the position, orientation, and size of the amodal bounding boxes projected back to all available cameras. The oriented bounding boxes are inspected from the top view too. The manual quality checking is performed on the scene level. Some label noise still might be included in the dataset even though we aimed to minimize it using human

validation. In this way, we selected sufficiently accurately labeled recordings, and most erroneous annotations were discarded.

In the case of the validation set, we hired manual annotators to label objects on the recorded sensor data. The human annotators used LiDAR and camera sensor data during the annotation phase to fit cuboids on any object of interest appearing on the camera images. For the cuboid sizes, annotators used default dimensions. If the default dimensions do not match the size of a given object in the point cloud or on the images, annotators adjusted the non-matching dimensions of the given cuboid based on their own decision. The manual labor also ensured that one cuboid axis is aligned with the object orientation and within 5 degrees of precision.

The manually or automatically annotated objects belonging to 14 classes are represented as 3D cuboids with some additional physical properties. Each labeled bounding box has a 3D center point, 3D extent (length along the horizontal x-axis, width along the vertical y-axis, height along the z-axis), orientation (represented as a quaternion), relative velocity, and a unique track ID. Furthermore, we provide 2D bounding boxes utilizing an FCOS (Tian et al., 2019) detector. The 2D-3D annotations are associated using the Hungarian algorithm (Kuhn, 1955) for allowing the utilization of 2D-3D consistency or semi-pseudo-labeling (Matuszka & Kozma, 2022). The resulting dataset was anonymized using Dashcam-Cleaner[1].

3.4 Dataset Analysis

The dataset includes 26 583 annotated frames with sensor data from multiple modalities, split into 21 402 train and 5 181 validation frames (80/20 train/val split). The scenes were recorded in diverse weather and environmental conditions. See Table 3 for the data distribution.

The dataset contains more than 425k objects organized into 14 categories. See the class distribution in Figure 4. The distance distribution of the annotated objects is visualized in Figure 5. About 24% of the cuboids are beyond 75 m, Argoverse2 has about 14%, Waymo, nuScenes, and ONCE have less than 1%. This property enables the training of long-range perception systems with the help of our dataset.

Several additional statistics of the generated dataset are described by Table 4, 8, Figure 6 and 7. The average cuboid dimensions for distinguished classes help to better understand how good the cuboids are per class. The number of average cuboids per environment indicates how crowded the scenes are. The percentage of empty boxes beyond 50 m and 75 m after the annotation process is 4.2% and 5.4%, respectively, as opposed to the conventional benchmarks where almost 50% of objects beyond 50 m contain zero LiDAR points (Gupta et al., 2023).

4 Experiments

We trained several 3D object detection baselines on our dataset using publicly available models. In order to utilize annotations located in distant areas, we defined the target grid as $[-204.8, 204.8]$ m in longitudinal and $[-25.6, 25.6]$ m in lateral directions. We mapped the 14 classes included in the dataset into four categories (car, truck/bus, motorcycle, pedestrian) and evaluated the model performance using the all-point and 11-point interpolated Average Precision (AP) metrics (Everingham et al., 2010) in Bird's-Eye-View (BEV) space in a class agnostic manner. The Hungarian method (Kuhn, 1955) is used for associating ground truth and predictions with a 0.3 IoU threshold. We selected a small IoU value for the association threshold to handle displacement errors which are especially frequent in distant regions in BEV. Furthermore, the Average Orientation Similarity (AOS) (Geiger et al., 2012a) metric is utilized for evaluating the performance of the models in terms of orientation prediction.

4.1 Baseline models

Our baseline models are based on VoxelNet (Zhou & Tuzel, 2018), BEVDepth (Li et al., 2022), and BEVFusion (Liu et al., 2022) for LiDAR, camera, and multimodal models. Since BEVFusion does not use radar sensors, we designed a simple solution for LiDAR-radar fusion. Namely, we treated

[1] https://github.com/tfaehse/DashcamCleaner

Modalities	Highway	Urban	Night	Rain	Highway	Urban	Night	Rain	AOS
LiDAR	0.755	0.608	0.746	0.477	0.731	0.600	0.727	0.480	0.594
LiDAR, camera	0.772	0.656	0.786	0.459	0.748	0.641	0.761	0.468	0.830
LiDAR, radar	0.769	0.617	0.732	0.454	0.745	0.607	0.748	0.465	0.509
LiDAR, radar, camera	0.762	0.644	0.730	0.423	0.740	0.630	0.708	0.435	0.851

Table 5: Comparison of baseline models. First group: all-point AP metric, second group: 11-point interpolation AP metric, third group: AOS metric averaged over val set.

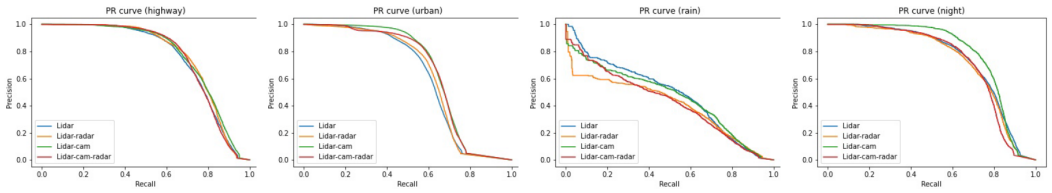


Figure 2: PR curves of baseline models.

the radar point cloud as a regular LiDAR point cloud. After a point cloud merging step, data from different modalities can be processed by VoxelNet as if it would be a regular LiDAR point cloud.

VoxelNet has the capability to operate on the point cloud directly and consists of three main parts. The Voxel Feature Encoder (VFE) is responsible for encoding raw point clouds at the individual voxel level. VoxelNet utilizes stacked VFE layers and their output is further processed by a middle convolutional neural network (CNN) to aggregate voxel-wise features. The final component performing the 3D object detection is the region proposal network (Ren et al., 2015).

BEVDepth is a camera-only 3D object detection network that provides reliable depth estimation. The main observation of the authors is that recent camera-only 3D object detection solutions utilizing pixel-wise depth estimation generate suboptimal results due to inadequate depth estimation. Therefore, explicit depth supervision encoding intrinsic and extrinsic parameters is utilized. In addition, a depth correction subnetwork is introduced using sparse depth data from a LiDAR point cloud to provide supervision for the depth estimation network.

The main contribution of BEVFusion is the utilization of the BEV space as the unified representation for camera and LiDAR sensor fusion. The image backbone proposed by BEVFusion explicitly predicts a discrete depth distribution for each image pixel, similar to BEVDepth (without the depth correction subnetwork). Then, a BEV pooling operator is applied on the 3D feature point cloud which is later flattened along the z-axis to get a feature map in BEV. The point cloud produced by a LiDAR is processed the same way as in the case of VoxelNet and the two BEV feature maps are fused by a CNN. Finally, the detection heads are attached to the output of the fusion subnetwork.

4.2 Implementation details

The LiDAR components of the baseline models use HardSimpleVFE (Yan et al., 2018) as the Voxel Feature Encoder and SparseEncoder (Yan et al., 2018) as the middle encoder CNN. The image components adopt Lift-Splat-Shoot (Philion & Fidler, 2020) as an image encoder with a ResNet-50 backbone followed by a Feature Pyramid Network (Lin et al., 2017) for leveraging multi-scale features. An additional depth correction network is also part of the image stream, inspired by BEVDepth. In the case of multimodal models, features from different modalities are fused using a simple fusion subnetwork consisting of convolution and Squeeze-and-Excitation (Hu et al., 2018) blocks. Finally, a CenterPoint (Yin et al., 2021) head is responsible for detecting objects from the BEV features both in unimodal and multimodal cases.

Since our goal is not to develop state-of-the-art models in this work but to facilitate multimodal object detection research, we used the hyperparameters provided by BEVDepth [2] without any heavy

[2] https://github.com/Megvii-BaseDetection/BEVDepth

Modalities	Highway	Urban	Night	Rain	Highway	Urban	Night	Rain	AOS
LiDAR	0.526	0.470	0.464	0.617	0.521	0.472	0.477	0.612	0.864
LiDAR, radar	0.522	0.525	0.550	0.652	0.522	0.519	0.546	0.644	0.868

Table 6: Comparison of baseline models in the distant region (>75m). First group: all-point AP metric, second group: 11-point interpolation AP metric, third group: AOS metric averaged over the validation set.

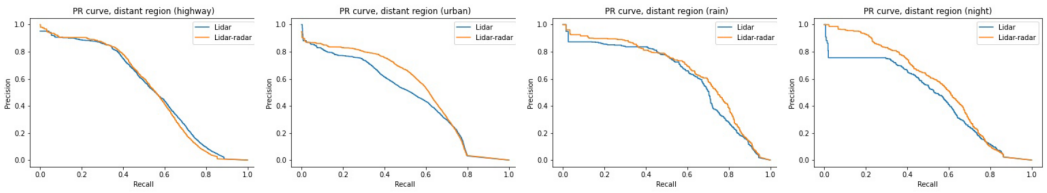


Figure 3: PR curves of baseline models in distant region (>75m).

parameter tuning. We adapted the grid resolution to enable long-range detection and trained the models for 16k iterations (3 epochs) using batch size 4 with a learning rate of $6.25e^{-5}$ using flip, rotation, and scale augmentations in the BEV feature space. We used an NVIDIA A100 TensorCore GPU for neural network training. The models are implemented using mmdetection3d[3] and will be made publicly available.

4.3 Experimental results

The performance comparison of the baseline models on different metrics is described in Table 5. Since the literature has several examples (Qian et al., 2022; Liu et al., 2022) of the superiority of LiDAR-only unimodal solutions over camera-only models, we did not train a camera-only baseline. As the table describes, every multimodal model overperforms the LiDAR-only baseline in highway and urban environments in non-adverse weather and time. The additional sensor signals significantly increase detection performance in the dense urban environment. However, the unimodal baseline performs best in heavy rain where one would think a radar signal should help to increase performance. This phenomenon suggests that more sophisticated radar fusion techniques might be beneficial for enhancing multimodal models.

Cameras play a crucial role in terms of orientation prediction. The models without RGB images struggle to consistently keep the orientation, especially in the case of large vehicles. This flickering effect is less visible for models using camera sensors. The model using all modalities performs best on the AOS metric.

Surprisingly, the model using LiDAR + camera modalities overperforms all other models in the night and urban environments by a large margin. We investigated the learning curves and found that increasing the number of training steps can help to enhance performance further. To validate our hypothesis, we trained our models for 5 additional epochs. Unfortunately, models using camera sensors became unstable after the third epoch causing an explosion in the depth loss. Table 7 describes the result of the longer training process using the 11-point interpolation AP metric. A solid improvement can be seen in all environments, especially on the rainy validation set ($+8.8$/$+5.2$ AP for LiDAR and LiDAR+radar models, respectively). This can be explained by the fact that the detection heatmaps became sharper after longer training. Blurry heatmaps were responsible for lower AP metrics in the case of the first group of baseline models. The blurring effect on the heatmap was very visible around the ego car in the case of heavy rain due to LiDAR reflections from raindrops.

In order to validate the long-range perception capabilities of the baseline models, we benchmarked the longer-trained models on distant object detection. Detections and ground truth were filtered out where the distance from the ego car is less than 75 meters. The results are summarized in Table 6

[3]https://github.com/open-mmlab/mmdetection3d

Modalities	Highway	Urban	Night	Rain
LiDAR	0.757 (+2.6)	0.630 (+3.0)	0.754 (+2.7)	0.568 (+8.8)
LiDAR, radar	0.741 (-0.4)	0.638 (+3.1)	0.766 (+2.2)	0.517 (+5.2)

Table 7: Effects of longer training.

(see Figure 10 for qualitative examples). Both models perform similarly in the highway environment without any significant difference in performance. However, the model with additional radar signals significantly overperforms the LiDAR-only baseline in all other environments. The fact, that radar sensors provide reliable and accurate signals for perceiving objects in distant areas even in adverse weather can be leveraged for boosting 3D object detector performance, as can be seen in Figure 3. A similar effect can be observed in a dense urban environment where radar signals are utilized by the multimodal baseline and resulted in a significant performance increase in long-range perception ($+5.5$/$+4.7$ all-point / 11-point interpolation AP).

The training results demonstrate that our dataset can serve as a base for multimodal long-range perception neural network training. Advanced evaluation techniques such as test-time augmentation or model ensembling could lead to further improvements. However, none of them were applied during the evaluation method. Table 7 suggests that further improvements in sensor fusion methods are needed for fully leveraging each modality and our naive method provides a suboptimal solution (especially in the case of heavy rain). Nevertheless, we hope the research community will find our dataset useful, can build on our baselines, and will significantly improve its performance.

5 Additional Proposed Tasks

We propose additional tasks benefiting from our dataset besides 3D object detection. Since unique track IDs are provided, end-to-end long-range multiple object tracking models can also be trained using the dataset. MOTA and MOTP metrics can be used for evaluating model performance.

Another proposed task is motion prediction. The egomotion is included in the dataset and the trajectories of exo-objects can be computed using the unique track IDs. We propose a specific case of motion prediction, namely lead car prediction, which is essential for autonomous driving functions such as Automatic Emergency Braking or Adaptive Cruise Control. The lead cars can be determined by the intersection of egomotion and the trajectories of exo-objects. The proposed task is to detect and predict the current and future lead cars. The model performance can be measured using precision and recall metrics.

The dataset also includes high-quality GNSS-INS sensory data, thus enabling the training and benchmarking of various odometry algorithms. Finally, the dataset can be used for contrastive representation learning. A similar representation can be learned for different sensor modalities corresponding to the same frame in a self-supervised manner.

6 Conclusion

In this paper, we present a multimodal dataset for robust autonomous driving with long-range perception. Our diverse dataset recorded in three countries on two continents includes sensor data from LiDAR, radars, and cameras providing redundant 360-degree sensor coverage. The dataset contains a large number of annotated objects in distant areas, allowing the development of multimodal long-range perception neural networks. In addition, we developed several unimodal and multimodal baseline models and compared their performance on the proposed dataset based on different criteria. We showed that our dataset can be used for training multimodal long-range perception neural networks leveraging the advantages of the recorded sensor modalities. In the future, we aim to extend our collected dataset with additional environmental and weather conditions. Furthermore, we will conduct more in-depth experiments regarding sensor fusion for multimodal neural networks. We seek to facilitate research in multimodal sensor fusion and robust long-range perception systems by releasing our dataset.

REFERENCES

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11682–11692, 2020.

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*, pp. 11621–11631, 2020.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8748–8757, 2019.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012a.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012b.

Shubham Gupta, Jeet Kanjani, Mengtian Li, Francesco Ferroni, James Hays, Deva Ramanan, and Shu Kong. Far3det: Towards far-field 3d detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 692–701, 2023.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 954–960, 2018.

Adrian Kaehler and Gary Bradski. *Learning OpenCV 3: computer vision in C++ with the OpenCV library*. " O'Reilly Media, Inc.", 2016.

R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. `https://level-5.global/level5/data/`, 2019.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv*, 2022.

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.

Tamás Matuszka and Dániel Kozma. A novel neural network training method for autonomous driving using semi-pseudo-labels and 3d data augmentations. In *Intelligent Information and Database Systems: 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28–30, 2022, Proceedings, Part II*, pp. 216–229. Springer, 2022.

Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9552–9557. IEEE, 2019.

Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A* 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2267–2273. IEEE, 2020.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pp. 194–210. Springer, 2020.

Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: a survey. *Pattern Recognition*, pp. 108796, 2022.

Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17021–17030, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–7. IEEE, 2021.

Richard A Snay and Tomás Soler. Modern terrestrial reference systems (part 1). *Professional Surveyor*, 19(10):32–33, 1999.

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*, pp. 2446–2454, 2020.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

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*, pp. 11784–11793, 2021.

Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.

A APPENDIX

A.1 DATASET STATISTICS

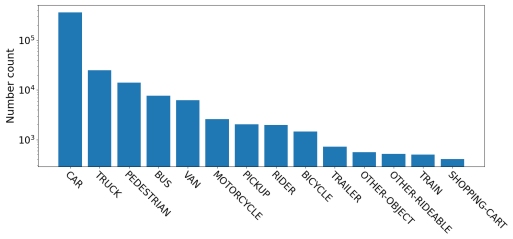


Figure 4: Class distribution.

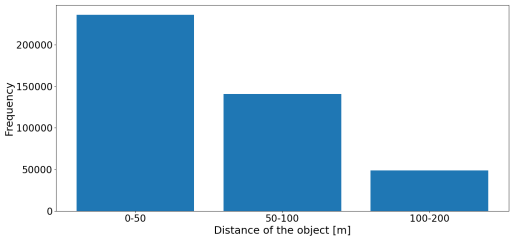
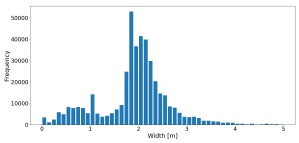
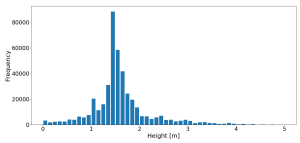


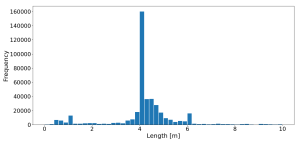
Figure 5: Distance distribution of annotated objects.



(a) Width distribution.



(b) Height distribution.



(c) Length distribution.

Figure 6: Dimension distribution of the annotated objects.

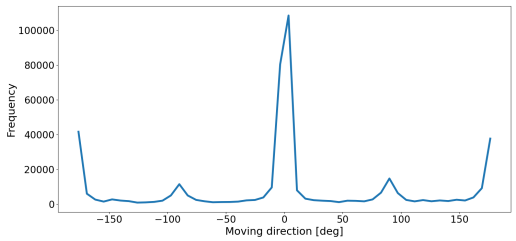


Figure 7: Orientation distribution of annotated objects.

(a) Mean density of scenes.

	Cuboids/scene
Urban	30
Highway	7
Night	14
Rain	6

(b) Points in cuboids.

	Mean	Std
0-50 m	1209	2378
50-75 m	91	126
75+ m	25	33

Table 8: Additional statistics of the annotations.

A.2 SAMPLES AND QUALITATIVE RESULTS

Figure 8: Example motorbike GT created by automatic annotation.

Figure 9: Example ground truth annotations.

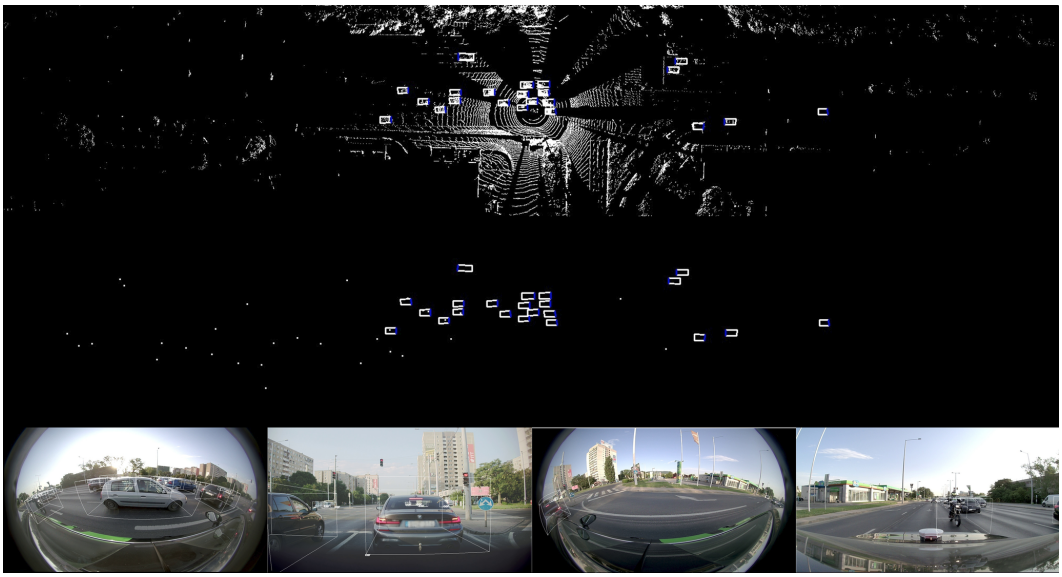


Figure 10: Qualitative results: detections of the LiDAR+radar baseline model. Top row: detections on LiDAR point cloud. Middle row: detections on radar targets, bottom row (from left to right): detections on left, front, right, and back cameras.