MAN TruckScenes: A multimodal dataset for autonomous trucking in diverse conditions

Felix Fent[1]   **Fabian Kuttenreich**[2]*   **Florian Ruch**[2]   **Farija Rizwin**[2]

Stefan Juergens[2]   **Lorenz Lechermann**[2]   **Christian Nissler**[2]   **Andrea Perl**[2]

Ulrich Voll[2]   **Min Yan**[2]   **Markus Lienkamp**[1]

[1]Technical University of Munich
School of Engineering & Design
Institute of Automotive Technology

[2]MAN Truck & Bus SE
`truckscenes@man.eu`

Abstract

Autonomous trucking is a promising technology that can greatly impact modern logistics and the environment. Ensuring its safety on public roads is one of the main duties that requires an accurate perception of the environment. To achieve this, machine learning methods rely on large datasets, but to this day, no such datasets are available for autonomous trucks. In this work, we present MAN TruckScenes, the first multimodal dataset for autonomous trucking. MAN TruckScenes allows the research community to come into contact with truck-specific challenges, such as trailer occlusions, novel sensor perspectives, and terminal environments for the first time. It comprises more than 740 scenes of 20 s each within a multitude of different environmental conditions. The sensor set includes 4 cameras, 6 lidar, 6 radar sensors, 2 IMUs, and a high-precision GNSS. The dataset's 3D bounding boxes were manually annotated and carefully reviewed to achieve a high quality standard. Bounding boxes are available for 27 object classes, 15 attributes, and a range of more than 230 m. The scenes are tagged according to 34 distinct scene tags, and all objects are tracked throughout the scene to promote a wide range of applications. Additionally, MAN TruckScenes is the first dataset to provide 4D radar data with 360° coverage and is thereby the largest radar dataset with annotated 3D bounding boxes. Finally, we provide extensive dataset analysis and baseline results. The dataset, development kit, and more are available online.

1 Introduction

Autonomous trucking has the potential to fundamentally change today's traffic by increasing safety on public roads, reducing logistics costs, and counteracting the shortage of drivers [2, 20]. However, the safe and reliable operation of autonomous trucks depends on an accurate perception of the surroundings. To achieve this, modern self-driving vehicles rely on machine learning algorithms to detect, track, and predict surrounding objects. However, the use of machine learning methods also drives the need for large-scale datasets.

While numerous datasets exist for autonomous passenger cars [15], datasets for autonomous trucks are missing. However, heavy-duty vehicles have their unique challenges. Large vehicles require

---

*corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

(a) terminal	(b) rain	(c) snow
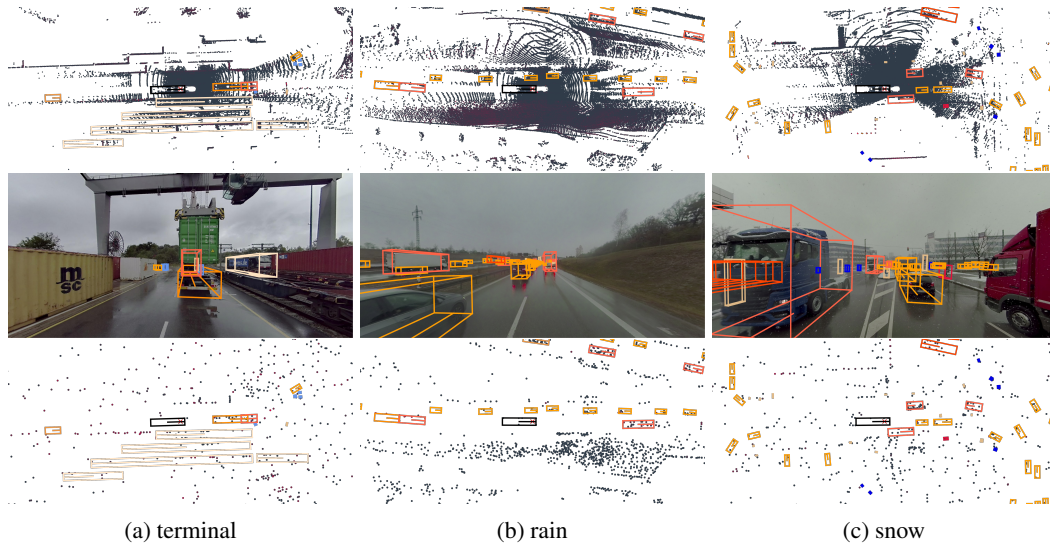
Figure 1: Exemplary selection of a terminal, rain, and snow scene of the MAN TruckScenes dataset. The top row shows the fused lidar point cloud, the center row shows images of the front left camera, and the fused radar point cloud is shown at the bottom.

different sensor mounting positions and rely on multiple sensors to cover the entire surrounding area. Moreover, trucks have to contend with occlusions from their own vehicle that change dynamically due to a movable truck-trailer combination and are affected by relative movements between their chassis and cabin. Furthermore, long-haul trucks operate in inherently different surroundings, such as logistics or container terminals and their functionality has to be ensured under all environmental conditions to maintain a functioning logistics system. Therefore, a dedicated truck dataset is needed to develop reliable perception solutions for self-driving trucks under all conditions.

To address this research gap and accelerate the development of self-driving trucks, we present MAN TruckScenes, the first large-scale dataset for autonomous trucking. Our multimodal dataset comprises data from a state-of-the-art sensor suite, including multiple high-resolution cameras, lidar, and radar sensors to provide full coverage of the surrounding area. Especially, the inclusion of six 4D radar sensors makes it the largest radar dataset available. In addition, the data of a high-precision GNSS and two IMU units are included to support a multitude of applications.

Moreover, most existing AV datasets are restricted to a particular operational design domain, whereas the MAN TruckScenes dataset covers a large geographical area, three seasons, nighttime driving, and numerous different weather conditions, including fog, rain, and snow. Furthermore, the dataset includes scenes in logistics terminals and recordings of vehicles with high relative velocities on the German Autobahn, making it the first dataset to promote the unique challenges of long-haul trucks.

In order to support the development of deep learning methods, MAN TruckScenes provides high-quality annotations for 747 scenes. These annotations consist of manually annotated 3D bounding boxes with unique instance identifiers for object tracking, standardized scene tags, and attribute labels to indicate the object's state. The dataset annotation was subject to a multi-stage labeling and quality assurance process of specially trained labelers to ensure accurate and consistent labeling.

The dataset will be published alongside a development kit, evaluation code, taxonomy, and detailed annotation instructions. This ensures transparency and reproducibility and simplifies the use of the dataset. Furthermore, we build on the established nuScenes [3] data format to ensure easy integration and code compatibility. Finally, MAN TruckScenes is published under the CC BY-NC-SA 4.0 license to accelerate research on autonomous trucks and shape the future of logistics.

2

Table 1: Comparison of publicly available perception datasets for autonomous driving. The coverage represents the geographical coverage calculated according to [22] and range refers to the 99.9th percentile of all bounding box distances [1]. Scenes are temporal consistent sequences and samples are annotated keyframes, following the nuScenes [3] notation. Vehicle refers to the recording vehicle.

	KITTI [11]	nuScenes [3]	Waymo [22]	Argo2 [23]	ZOD [1]	aiMotive [17]	VoD [18]	Ours
Scenes	22	1000	1150	1000	1473	176	21	747
Sample	1.5 k	40 k	230 k	150 k	1.5 k	27 k	9 k	30 k
Duration	1.5 h	5.5 h	6.4 h	4.2 h	8.2 h	0.7 h	0.2 h	4.2 h
Coverage	-	4 km^2	76 km^2	17 km^2	26 km^2	180 km^2	2 km^2	100 km^2
Camera	4	6	5	9	1	4	1	4
Lidar	1	1	5	2	3	1	1	6
Radar	0	5	0	0	0	2	1	6
GNSS	1	1	0	0	2	1	1	1
IMU	1	1	0	0	2	1	0	2
Map	no	yes	no	yes	no	no	no	no
Range	91 m	141 m	80 m	214 m	245 m	228 m	26 m	226 m
Classes	3	23	4	30	15	14	13	27
Vehicle	car	car	car	car	car	car	car	truck

2 Related Work

The advances in autonomous driving have largely been driven by the release of public datasets. Over the course of the last decade, we have seen a trend towards high-quality, large-scale, and more diverse multimodal datasets that enabled innovation in the autonomous driving domain.

The KITTI [11] dataset, released in 2012, is one of the most influential autonomous driving datasets to date. It provides 22 scenes of annotated camera and lidar data in combination with high-precision GNSS and IMU data. However, the dataset's extent and diversity are limited, radar and map data are not provided, and driving data under severe weather conditions are not included.

The nuScenes [3] dataset had a great impact as one of the first large-scale, multimodal datasets for autonomous driving. Next to camera and lidar data, it also includes radar and map data for 1000 annotated scenes. It provides data from two different cities, annotates 23 different object classes, and introduces unlabeled intermediate sensor data sweeps. In addition to the small geographical coverage and limited coverage of severe weather conditions, the nuScenes dataset has been criticized for its sparse radar data [6, 19].

The Waymo Open [22] dataset is one of the largest annotated datasets for autonomous driving with a focus on scalability. It provides annotated camera and lidar data for 1150 scenes, covers a geographical area of 76 km^2, and provides a rich ecosystem of different related datasets. However, it does not include radar, GNSS, or map data and has a limited annotation range of 80 m with only four annotation classes.

The Argoverse 2 [23] dataset takes a different approach with the provision of long-range annotations, comprehensive map data, and extensive object taxonomy. Nevertheless, it does not provide radar, GNSS, or IMU data and covers a limited geographical area.

The Zenseact Open Dataset (ZOD) [1] makes a different set of trade-offs. It provides 1473 annotated scenes from six different countries and facilitates long-range perception with an annotation range of 245 m. Despite its provision of two GNSS and IMU units, the ZOD's sensor setup, as well as its number of annotated samples, is limited.

The aiMotive [17] dataset focuses on the collection of diverse driving scenes under severe weather conditions. It provides long-range annotations and data from a multimodal sensor setup for three distinct areas. However, the dataset's extent is limited to a duration of 0.7 h and the two radar sensors do not cover a 360° field of view (FoV).

3

Table 2: Sensor specifications of the MAN TruckScenes setup.

Sensor	Details
Camera	4× Sekonix SF3324, RGB, 10 Hz, 1928 × 1208, 120° × 73° FoV
Lidar	2× Hesai Pandar64, 10 Hz, 64 layer, 360° × 40° FoV, 200 m@10 %
	4× Ouster OS0, 10 Hz, 64 layer, 360° × 90° FoV, 35 m@10 %
Radar	6× Continental ARS 548 RDI, 20 Hz, 76 GHz, 100° × 28°
GNSS	1× GeneSys ADMA-G-PRO+, 100 Hz, 0.01 m pos., 0.015° heading
IMU	2× Xsens MTi-680G-SK, 100 Hz, 9 DoF
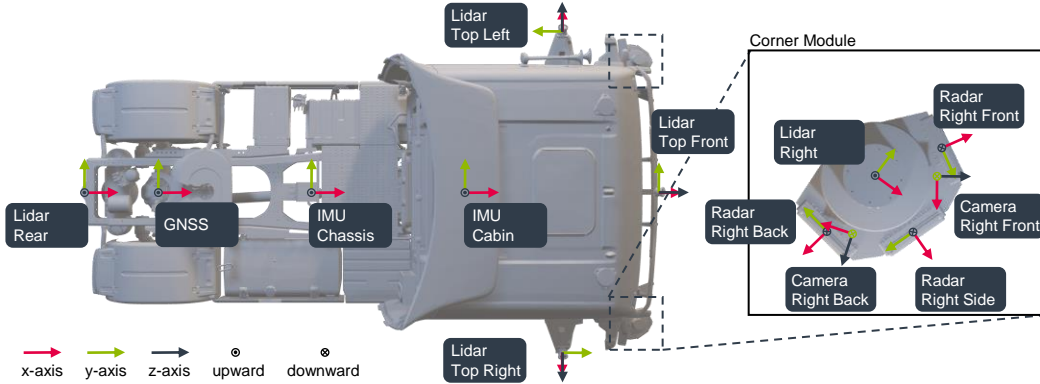
Figure 2: Placement of the sensors and their corresponding coordinate systems from a top view perspective. The right sensor module is shown in the detailed drawing, the left module is identical but mirrored and the radar sensors are flipped around the x-axis. The top-mounted lidar sensors (at the sides and the front) are tilted downward. The GNSS uses a virtual coordinate system similar to the vehicle frame but is located next to the chassis IMU.

The View-of-Delft [18] dataset is one of the first large-scale datasets with annotated 4D radar data and promotes the development of radar-based perception methods for vulnerable road users (VRUs) in city environments. Nevertheless, the geographical coverage of the dataset is limited, it does not include severe weather conditions, and only provides data from a single front-facing radar sensor.

While there are numerous other datasets focusing on various different aspects [15], all of the aforementioned datasets are limited to passenger cars, as shown in Table 1. To the best of our knowledge, there are only two datasets that include truck data. One is the TuSimple [25] dataset, which consists of 6408 images from a single front camera and provides annotations for lane markings only. The other is the SurMine [21] dataset, which is a proprietary dataset that only includes 2D bounding box annotations for 10665 camera images recorded in a mining facility. Hence, there are no large-scale or multimodal perception datasets for autonomous trucking.

3 Dataset

MAN TruckScenes aims to close this research gap by providing the first large-scale multimodal dataset for autonomous trucking. It consists of 747 scenes from typical long-haul truck environments and provides multimodal data with long-range annotations based on the nuScenes [3] format.

3.1 Sensor Setup

The sensor suite consists of a multimodal sensor setup with four cameras, six lidar, and six radar sensors. Additionally, the dataset provides high-precision RTK-GNSS data and measurements of two IMU units. Detailed information on the sensor specifications can be found in Table 2 and the sensor positions are shown in Figure 2.

The main perception sensors are arranged in two sensor modules, one on either side of the vehicle, to maximize spatial coverage and minimize relative sensor movements. Each sensor module houses two

4

Sekonix cameras, one Hesai lidar, and three Continental radar sensors. Besides these two "corner modules", the vehicle is equipped with three Ouster lidar sensors mounted on the roof of the cabin, which are tilted downwards for blindspot coverage, and one additional Ouster lidar at the rear of the semi-trailer truck. The top mounted lidar sensors and the lidar sensors of the corner modules are positioned at a height of 3.2 m and 2.2 m, respectively. These elevated positions help to reduce occlusions, protect pedestrians, and prevent sensor damage. However, this sensor perspective marks a significant difference from conventional passenger car datasets.

Another unique feature of our dataset is the inclusion of six 4D radar sensors with a spatial coverage of nearly 360° (except from the occlusion by the ego vehicle's trailer). In contrast to conventional 3D radar sensors that operate in the range-azimuth plane, 4D radar sensors can resolve objects in both azimuth and elevation angle. Furthermore, our radar sensors provide on average 2600 points per sample, while conventional radar sensors (e.g. in the nuScenes dataset) provide only 200 data points. Therefore, MAN TruckScenes is the first dataset to provide 4D radar data with 360° coverage and is the largest radar dataset with annotated 3D bounding boxes.

To accurately capture the vehicle's state and position, the data of two IMUs and one RTK-GNSS are included. The GNSS unit uses a dual antenna setup to measure the heading angle of the vehicle and RTK correction data to achieve a position accuracy of up to 0.01 m. The two IMU units are mounted on the chassis and the cabin of the vehicle to get accurate measurements of the overall vehicle state on one side and measure relative movements between the chassis and the cabin on the other. This is important to compensate for sensor movements that are primarily mounted on the movable cabin and represent a special challenge of the truck case.

3.2 Sensor Synchronization

Sensor synchronization is an important topic for multimodal datasets to ensure temporal sensor data alignment. This includes not only sensor time synchronization, which ensures that all sensors are based on the same reference clock, but also triggering the sensors to achieve cross-modality consistency.

The sensor time synchronization is based on the Precision Time Protocol (PTP) in accordance with the IEEE/IEC 1588-2008 [12]. This procedure ensures the alignment of the individual sensor clocks with a high-precision Mobatime DTS 4160 PTP grandmaster clock. As a result, the individual sensors have a time deviation of less than 100 μs and are referenced to the global UTC time. However, this only ensures that all sensors are based on the same reference clock, but not that all measurements are taken at the exact same point in time.

To temporally align the actual sensor measurements, the sensors have to be triggered. For this purpose, the lidar sensors are defined as reference sensors and all other perception sensors are triggered based on them. In the first step, all cabin-mounted lidar sensors are phase-synchronized such that their lidar points are temporally consistent in a clockwise manner. Secondly, the four cameras are triggered at the point where the lidar sweep and the rolling shutter of the cameras align in the center of the image. Lastly, the radar sensors are synchronized with the corresponding lidar sensors but triggered with a small delay to one another to minimize interference between them. Besides that, all radar sensors are assigned to slightly different frequency bands to further minimize interference.

3.3 Sensor Calibration

While sensor synchronization ensures the temporal alignment of the sensor data, sensor calibration ensures the spatial alignment. The calibration aims to determine both the extrinsic and intrinsic sensor parameters to estimate both the sensor poses as well as the parameters of the sensor models.

The sensor calibration consists of four steps and includes an initial extrinsic calibration, a combined extrinsic and intrinsic calibration, an angular correction, and a final validation. The first step is a photogrammetric scan of the vehicle to determine the position and orientation of the sensors with respect to the vehicle frame. The vehicle frame is chosen to be the center of the rear axle projected onto the ground plan in accordance with ISO 8855:2011 [14]. For the main calibration, the vehicle is placed in a dedicated calibration hall equipped with specialized calibration targets. These targets are represented by differently oriented planes in 3D space equipped with unique identifiers. The actual calibration uses a plane-matching algorithm to jointly estimate the extrinsic parameters of the

5

camera and lidar sensors as well as the intrinsic camera parameters. Within this calibration procedure, the vehicle is moved back and forth to calibrate the sensors with respect to the vehicle frame. The third step uses a faraway calibration target aligned with the vehicle's longitudinal axis to correct remaining yaw angle errors and a dynamic landmark-based calibration to align the heading angle of the overall sensor setup. Finally, the calibration is validated on the data level by comparing point cloud alignments and projecting the point cloud data onto the image data. Moreover, an application-level validation is used to compare detections, odometry, and localization measurements to validate the calibration.

This procedure results in a precise calibration with pixel-level accuracy between camera and lidar sensors. However, the radar calibration solely relies on external measurements of the photogrammetric scan with a measurement tolerance of $0.05\,$mm. Therefore, angular errors of up to $0.03°$ and sensor internal misalignments cannot be excluded. Besides that, the odometry-based validation can only ensure a heading error of less than $0.1°$ and one has to be aware of increasing uncertainties within the camera intrinsic calibration towards the edges of the camera image.

3.4 Sensor Data

The provided sensor data was going through different processing steps to ensure compatibility with the nuScenes [3] format and enhance usability. The properties of the five different sensor data types are explained in the following.

The camera data of the four camera sensors undergo an undistortion process using a pinhole camera model and a Lanczos interpolation scheme with a 8×8 kernel and constant padding values. The resulting image is cropped to a 1980×943 pixel size and stored as a compressed JPEG [13] image. The Camera data is sampled at a frequency of $10\,$Hz in correspondence with the lidar.

The lidar data is provided as point clouds of a variable number of points represented by their $x, y,$ and z-coordinates in a cartesian sensor coordinate system, an intensity value, and an individual UNIX timestamp, allowing for point-wise motion compensation. The final point clouds are stored as binary (Marc Lehmann's LZF) compressed data in the point cloud data (pcd) format to save on disc space.

The radar data is also represented as point clouds where each point is defined by its $x, y,$ and z-coordinates, its relative radial velocity in $x, y,$ and z-direction as well as its radar cross section (rcs). In contrast to the lidar data, the radar data is stored as pcd files in a binary format due to its smaller file size.

The IMU data includes the current velocity and acceleration in $x, y,$ and z-direction as well as the angle and angular velocity (rate) in roll, pitch, and yaw. The data is stored in a JSON file. The GNSS data provides the vehicle's position in UTM-WGS84 coordinates mapped to cell U32 and orientation given as quaternion (qw, qx, qy, qz). The GNSS data off all timestamps is stored in a JSON file. IMU and GNSS data are sampled with a frequency of $100\,$Hz to provide the possibility for highly accurate motion estimation.

Besides that, the rear lidar is limited to an FoV of $180°$, which increases its detection range to $42\,$m at $10\,\%$ reflectivity, and the points of the top mounted lidar sensors are cut off for yaw angles beyond $\pm120°$. It is also worth mentioning that the sensor data is not just provided at the sample annotation frequency of $2\,$Hz but at their individual sensor captioning frequencies listed in Table 2. These unannotated intermediate frames are denoted as sweeps. It is also important to note that the radar sensors cannot guarantee a fixed sampling rate of $20\,$Hz but rather drop frames, if their internal data processing takes too long, which results in an average data rate of $19.6\,$Hz.

3.5 Scene Selection

The scene selection aims to select a diverse set of scenarios with challenging driving situations representing long-haul trucks' operational design domain. To meet this goal, 747 scenes with approximately $20\,$s each are manually selected from more than $25\,$h of measurement drives. The scenes include recordings from various different areas (e.g. highway, terminal, rural, city), three seasons of the year, challenging weather situations (e.g. rain, fog, snow), and difficult environmental conditions (e.g. nighttime, twilight). Furthermore, the dataset covers different driving maneuvers (e.g. overtaking, offloading), traffic scenarios, and observations of rare object classes (e.g. animals,

6
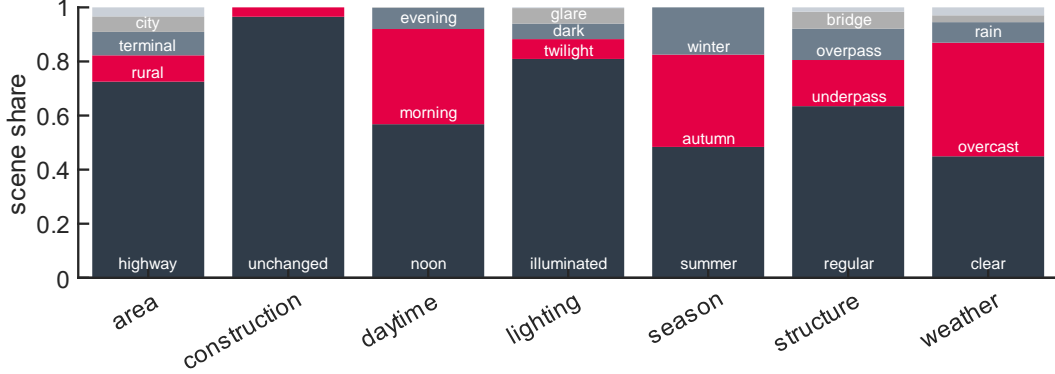
Figure 3: Distribution of the scene tags for all 747 dataset scenes.

emergency vehicles). As a result, the dataset includes rainy nighttime drives that cause mirroring effects in the lidar point cloud, recordings in a logistics terminal with a lot of metal shipping containers which lead to multipath reflections in the radar point cloud, and tunnel passages with strong illumination changes that effect the camera images. Therefore, the dataset provides challenging conditions for all sensor modalities to promote research in the area of robust perception. The distribution of the different scene tags is shown in Figure 3

3.6 Data Annotation

The data of the selected scenes is manually annotated at 2 Hz based on a fused and ego motion-compensated lidar point cloud aggregation. To ensure a high-quality standard, independent annotation and quality assurance companies have been commissioned. Within the labeling process, each sample undergoes a maximum of three consecutive annotation and quality assurance cycles until the quality target is met. Furthermore, randomly selected samples pass through a dual-control cycle to validate the annotation quality.

Annotations are made in the form of 3D bounding boxes defined by their center point (x, y, z), size (w, l, h), and orientation (qw, qx, qy, qz), given as quaternion. Each bounding box instance is classified according to 27 different object classes, using a hierarchical class structure, and in accordance with the nuScenes [3] data format. The distribution of the object categories is shown in Figure 4. In addition, every individual bounding box is labeled according to five attributes (with a total of 15 possible values) representing its visibility level or activity state. Furthermore, objects are tracked throughout the scene and assigned a unique and consistent tracking ID. Besides that, we label all scenes according to 34 distinct scene tags divided into seven different categories, including area, weather, and lighting conditions, as shown in Figure 3.

3.7 Dataset Splits

The dataset is split into a train, validation, and test set to ensure an independent evaluation. While these splits should be different from one another to test the generalization capabilities of a method, the data splits should also have similar characteristics to ensure a fair evaluation.

To address this conflict of objectives, we developed a method to find a Pareto-optimal solution for this multi-objective optimization problem (MOOP). For this purpose, the optimization uses the NSGA-II [4] genetic algorithm with random sampling, polynomial mutation [5], and simulated binary crossover [5] without duplicates. The optimization problem

$$\min_{s \in S}(f_1(s), f_2(s), ..., f_{12}(s))$$
$$|S_{\text{train}}| = 0.7|S|, \quad |S_{\text{test}}| = 0.2|S| \tag{1}$$

is given as a minimization problem with constraints, where S is the set of all scenes s defined by their scene properties (number of object classes, scene tags, sample timestamps, and ego positions) with cardinality $|S|$. The objectives $f_1(s)$ to $f_8(s)$ are the minimization of the deviations of the discrete

7
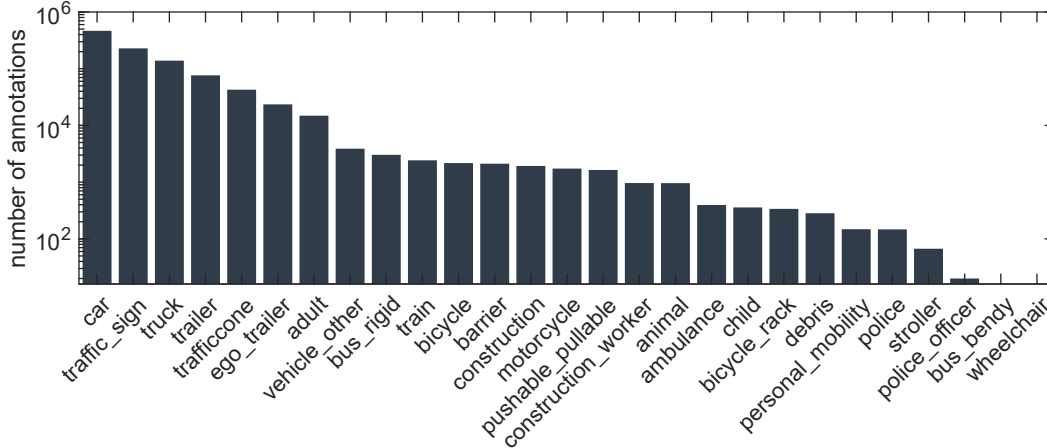
Figure 4: Number of annotated 3D bounding boxes for all 27 object classes across all dataset splits.

distributions of annotations across dataset splits, which are chosen to be the class distribution and the distributions of the seven individual scene tag categories. The objectives $f_9(s)$ and $f_{10}(s)$ aim to maximize the intra-split standard deviation of temporal (sample timestamps) and spatial (ego vehicle positions) components of the scenes. Finally, $f_{11}(s)$ and $f_{12}(s)$ aim to maximize the inter-split Kullback–Leibler (KL) divergence of the temporal and spatial components. The split sizes are set to 70 %, 10 %, and 20 % of all scenes for the train, validation, and test splits.

3.8 Privacy

Compliance with data protection measures is a top priority, which is why the whole dataset is anonymized. The anonymization includes not only the blurring of faces and license plates in the image data but also the anonymization of timestamp information. Nevertheless, the temporal consistency is still guaranteed for all samples and sensor data. As shown by Alibeigi et al. [1], the chosen image anonymization should not affect the downstream perception tasks.

4   Tasks

MAN TruckScenes supports a multitude of different perception tasks through its multimodal nature, sequential data structure, and rich annotations. These tasks include object detection, tracking, prediction, and localization, even if we want to put special emphasis on 3D object detection.

The detection task requires detecting 3D bounding boxes of 12 different object classes which are a subset of the original 27 annotation classes. These classes are selected based on the nuScenes [3] detection task and the insides gained during our labeling campaign. As a result, the evaluation excludes object classes that are not present in all data splits and combines subclasses with high inter-class confusion (seen during labeling) in accordance with our hierarchical class structure.

The evaluation is based on the nuScenes Detection Score (NDS), which is a weighted sum of the mean Average Precision (mAP) and five True Positive (TP) metrics [3]. In contrast to an Intersection over Union (IoU) based mAP, the NDS uses a distance-based mAP and averages the individual Average Precision (AP) values not only over the classes but also over four discrete distance thresholds. In addition to the mAP, the NDS takes five TP metrics into account, which are the Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). Further details on the NDS and its calculation can be found in [3].

In contrast to the NDS, our detection range is not limited to 50 m during evaluation but considers objects of up to 150 m around the vehicle. This should emphasize the development of long-range detection methods, which are crucial for the safe operation of autonomous vehicles on highways. Furthermore, we introduce the animal and traffic sign classes as two additional detection classes, leading to 12 distinct detection classes.

8

Table 3: Baseline results on the test split of the MAN TruckScenes dataset v1.0.

Method	Dataset	Modality	mAP	NDS	ATE	ASE	AOE	AVE	AAE
PETR [16]	MAN	camera	0.02	0.12	1.13	0.69	0.65	1.50	0.56
RadarGNN [10]	MAN	radar	0.07	0.11	0.89	0.81	1.13	8.00	0.57
CenterPoint [24]	MAN	lidar	0.27	0.41	0.41	0.35	0.28	2.73	0.20

5 Experiments

We provide baseline 3D object detection results for all three sensor modalities, which are shown in Table 3. The camera results are based on the PETR [16] model architecture with a pre-trained FCOS3D (V-99-eSE) backbone and an adjusted detection head to support 12-class classification. The detection distance is set to $\pm 150\,\mathrm{m}$ and the final image resolution is chosen to be 320×800 pixels. The radar baseline is set by RadarGNN [10], which was trained on fused, ego-motion compensated 4D radar point clouds with 6 aggregated sweeps. The node features consisted of the rcs values, timestamps, connectivity degrees, and the relative radial velocities of the radar points. Lastly, a CenterPoint [24] model was trained on fused ego-motion compensated lidar point clouds with 3 aggregated sweeps cropped to a $300\,\mathrm{m} \times 300\,\mathrm{m}$ grid with a voxel resolution of $(0.1\,\mathrm{m}, 0.1\,\mathrm{m}, 0.2\,\mathrm{m})$. The model used seven detection heads and was evaluated for 12 detection classes, according to our metric definition.

The results suggest that the detection quality of diverse object classes (like 'animal' or 'other vehicle') is insufficient and that the overall detection quality decreases significantly with increasing distance. Moreover, the results show a reduced detection quality in winter conditions and tunnels. Furthermore, it can be observed that the camera model provides insufficient results for minority classes and struggles with challenging weather and lighting conditions, whereas the radar-based detection model provides insufficient results for all non-metallic object classes. Moreover, it can be observed that the radar performance is negatively affected by reflections within tunnels or container terminals. The lidar performance, on the other hand, is reduced under rainy or foggy conditions but also struggles within tunnel scenarios and seems to be more sensitive to the number of available training samples. In general, the results show that a more robust long-range perception is required for safe autonomous trucking.

6 Conclusion

In this work, we presented MAN TruckScenes, the first dataset for autonomous trucking. It is a large-scale multimodal dataset for the development of autonomous truck perception applications. The dataset consists a diverse set of scenes recorded in a multitude of different environmental conditions and provides 360° coverage of all sensor modalities, including 4D radar data. It contains carefully reviewed 3D bounding boxes, tracking information, and rich annotations for all objects within more than 230 m range. Furthermore, we introduced a generic method for compiling meaningful and balanced data splits. We provide baseline results for 3D object detection and promote research in all areas of perception, tracking, and prediction. For this purpose, we will maintain a public leaderboard and provide annotation instructions, taxonomy specifications, and a development kit. With this dataset, we aim to accelerate the development of autonomous trucking.

7 Limitations

The dataset is limited to measurements on public roads and logistic terminals in Germany recorded on a single autonomous test truck. The data distribution represents the operational design domain of a long-haul truck and is not representative for other distribution haulage. The dataset is manually annotated and, therefore, subject to human annotation errors even if our extensive quality assurance process seeks to minimize them. The extrinsic sensor calibration can be affected by cabin movements, the IMUs are exposed to vehicle vibrations, and the ego vehicle position is subject to GNSS inaccuracies even with RTK correction.

9

8   Societal Impact

Despite the potential benefits of autonomous trucks the societal impact of the technology must be considered. Since autonomous trucks operate on public roads, their safety is a major concern that must be assured and cannot be compromised for economic reasons. While safety assurance is not just the key deployment factor, it is also the main driver for public acceptance and trust [7]. Therefore, regulatory authorities have to establish standards that ensure the safety of autonomous trucking and allow cross-border operation. Besides that, the impact on labor must be taken into account. While some studies expect job losses through autonomous vehicles [7], others argue that trained personnel will still be needed in the overall logistics process to supervise the system or carry out manual tasks [9]. Recent studies show that advanced automation can even help to improve working conditions and therefore, counteract the shortage of drivers and minimize turnover rates [8]. Ultimately, regulatory measures must be put in place to ensure that the interests of the general public are protected.

Acknowledgment

MAN TruckScenes would not have been possible without the support of many. For this reason, we would like to thank the whole MAN team for their dedicated and passionate work. We would like to thank the MAN workshop and integration team for their tireless support and the MAN management and project team for the strong promotion of this project. Last but not least, we want to thank our TRATON and Scania colleagues for their friendly support and expertise. The dataset annotation and quality assurance were handled by b-plus. This work was partially funded by the government-supported ATLAS-L4 project with grand no. 19A21048I.

References

[1] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson. Zenseact Open Dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023.

[2] G. Bray and D. Cebon. Operational speed strategy opportunities for autonomous trucking on highways. *Transportation Research Part A: Policy and Practice*, 158:75–94, Apr. 2022.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, Apr. 2002.

[5] K. Deb, K. Sindhya, and T. Okabe. Self-adaptive simulated binary crossover for real-parameter optimization. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, GECCO07. ACM, July 2007.

[6] F. Engels, P. Heidenreich, M. Wintermantel, L. Stacker, M. Al Kadi, and A. M. Zoubir. Automotive Radar Signal Processing: Research Directions and Practical Challenges. *IEEE Journal of Selected Topics in Signal Processing*, 15(4):865–878, June 2021.

[7] J. Engström, R. Bishop, S. E. Shladover, M. C. Murphy, L. O'Rourke, T. Voege, B. Denaro, R. Demato, and D. Demato. *Deployment of Automated Trucking: Challenges and Opportunities*, page 149–162. Springer International Publishing, June 2018.

[8] S. Escherle, E. Darlagiannis, and A. Sprung. Automated trucks and the future of logistics: A delphi-based scenario study. *Logistics Research*, 16(1):1–21, 2023.

[9] S. Escherle, A. Sprung, and K. Bengler. *How Will Automated Trucks Change the Processes and Roles in Hub-to-Hub Transport?*, page 51–69. Springer Nature Switzerland, 2023.

[10] F. Fent, P. Bauerschmidt, and M. Lienkamp. Radargnn: Transformation invariant graph neural network for radar-based perception. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2023.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, Aug. 2013.

10

[12] IEEE 1588. IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. Standard, IEEE The Institute of Electrical and Electronics Engineers, New York City, US, Nov. 2019.

[13] ISO 15444:2019. Information technology - JPEG 2000 image coding system. Standard, International Organization for Standardization, Geneva, CH, Oct. 2019.

[14] ISO 8855:2011. Road vehicles - Vehicle dynamics and road-holding ability. Standard, International Organization for Standardization, Geneva, CH, Nov. 2013.

[15] M. Liu, E. Yurtsever, J. Fossaert, X. Zhou, W. Zimmer, Y. Cui, B. L. Zagar, and A. C. Knoll. A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook. *IEEE Transactions on Intelligent Vehicles*, page 1–29, 2024.

[16] Y. Liu, T. Wang, X. Zhang, and J. Sun. *PETR: Position Embedding Transformation for Multi-view 3D Object Detection*, page 531–548. Springer Nature Switzerland, 2022.

[17] T. Matuszka, I. Barton, Á. Butykai, P. Hajas, D. Kiss, D. Kovács, S. Kunsági-Máté, P. Lengyel, G. Németh, L. Pető, D. Ribli, D. Szeghy, S. Vajna, and B. V. Varga. aimotive dataset: A multimodal dataset for robust autonomous driving with long-range perception. In *International Conference on Learning Representations 2023 Workshop on Scene Representations for Autonomous Driving*, 2023.

[18] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, Apr. 2022.

[19] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wohler. RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. IEEE, Nov. 2021.

[20] A. M. Schuster, S. Agrawal, N. Britt, D. Sperry, J. A. Van Fossen, S. Wang, E. A. Mack, J. Liberman, and S. R. Cotten. Will automated vehicles solve the truck driver shortages? Perspectives from the trucking industry. *Technology in Society*, 74:75–94, Aug. 2023.

[21] R. Song, Y. Ai, B. Tian, L. Chen, F. Zhu, and F. Yao. MSFANet: A Light Weight Object Detector Based on Context Aggregation and Attention Mechanism for Autonomous Mining Truck. *IEEE Transactions on Intelligent Vehicles*, 8(3):2285–2295, Mar. 2023.

[22] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020.

[23] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[24] T. Yin, X. Zhou, and P. Krahenbuhl. Center-based 3D Object Detection and Tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021.

[25] S. Yoo, H. S. Lee, H. Myeong, S. Yun, H. Park, J. Cho, and D. H. Kim. End-to-end lane marker detection via row-wise classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2020.

11

A Appendix

The appendix provides additional details on the MAN TruckScenes dataset and the conducted experiments. The sections are structured in accordance with the main body.

A.1 Sensor Synchronization

Sensor synchronization is especially important to ensure temporal consistency in multimodal datasets. As mentioned in Section 3.2, all cabin-mounted lidar sensors are phase-synchronized, such that their lidar points are temporally consistent in a clockwise manner. However, this approach leads to a temporal discontinuity at one point of a full cycle. This point is chosen to be in the back of the vehicle where the rear lidar is located. Furthermore, perfect consistency cannot be achieved for the top-mounted lidar sensors due to their tilted axis of rotation and constant rotational speed. Nevertheless, local deviations are small, such that the fused point clouds of a single sample have similar temporal properties as a single top-mounted lidar sensor. This temporal consistency of the lidar point clouds is shown in Figure A1.
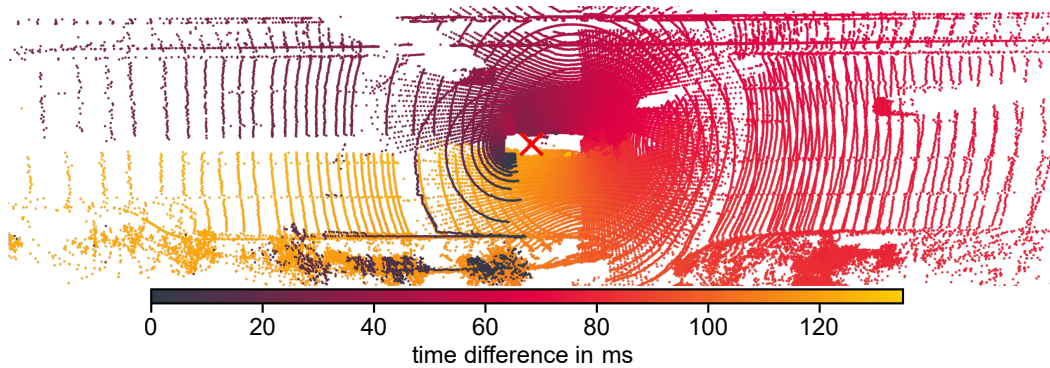
Figure A1: Lidar point clouds of all six lidar sensors color-coded by their time difference to the earliest recorded point in all six point clouds.

To comply with the nuScenes data format all annotations are made at a specific point in time sampled at a frequency of 2 Hz. The annotation time is chosen to be the point in time when all lidar scans align at the longitudinal axis in front of the vehicle. Therefore, the individual sensor timestamps have an offset to the sample (annotation) timestamp, as shown in Figure A2. To account for the differences in the recording time, the annotations are based on a fused and ego-motion compensated point cloud, as described in Section 3.6. However, this procedure can only compensate for ego-motion and not for target-motion. This can lead to inconsistencies, especially at the point of temporal discontinuity in the back of the vehicle.
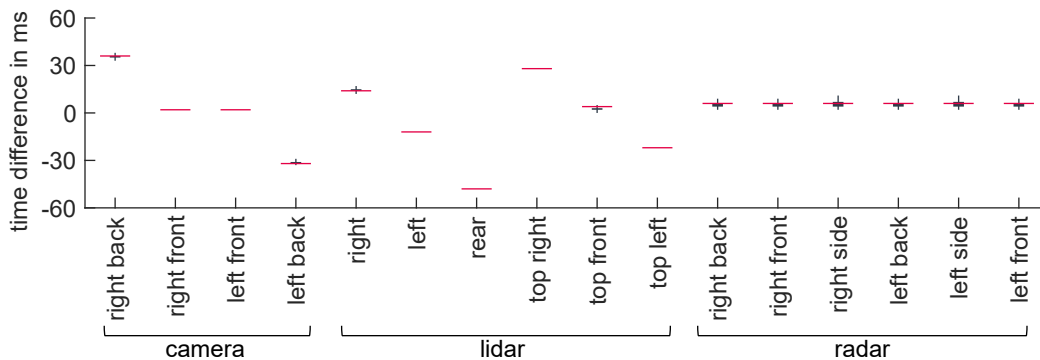
Figure A2: Differences of the sensor timestamps to the sample timestamp represented by their median offset (red lines), interquartile range (blue boxes), and whiskers (blue lines) across all samples.

12

A.2 Sensor Calibration

The quality of the sensor calibration is illustrated by a projection of the lidar points onto the camera images, as shown in Figure A3. It is important to note that a matching projection can only be achieved within a static scenario for multiple reasons. First, differences in the recording time caused by small deviations in the sensor synchronization (Figure A2) can lead to changes in the captured environment. Second, the rotating measurement principle of the lidar sensors leads to a temporal gradient (from left to right) in the recording of the lidar points (Figure A1). Third, the rolling shutter effect of the cameras causes a temporal gradient (from top to bottom) in the recording of the pixel rows. All these effects can lead to inconsistencies in the measurements of the different sensor modalities.

Figure A3: Projection of the left lidar sensor onto the front left camera image (left) and the right lidar sensor on the right front camera image (right). The color of the points indicates the relative distance.

A.3 Sensor Data

The provided sensor data is subject to multiple processing steps described in Section 3.4. Within this process, the camera data is mapped to a pinhole camera model to ensure its compatibility with the nuScenes data format and most existing perception methods. Therefore, the images of the wide-angle cameras are undistorted and cropped to remove black spots from the optics. The differences between the original and processed camera images can be seen in Figure A4.

Figure A4: Comparison of the original (left) and the undistorted (right) camera image.

A.4 Scene Selection

The scene selection aims to select diverse scenarios representative of a long-haul truck's operation, as described in Section 3.5. This selection primarily consists of highway, terminal, and rural scenes as well as scenarios from city and residential areas. To still be able to provide a diverse set of scenes, MAN TruckScenes covers a geographical area of $100\,\mathrm{km}^2$, measured by the union area of the 150 m-diluted ego-poses, as defined in [22]. The geographical locations of all included scenes are shown in Figure A5. While MAN TruckScenes does not provide map data, the included GNSS data enables the usage of open geographic databases (like OpenStreetMap) to utilize external map data, as shown in Figure A5.

A.5 Data Annotation

This section provides additional information on the included annotations of the dataset. As mentioned in Section 3.6, the annotations comprise scene tags, 3D bounding boxes, tracking IDs, and object

13

Figure A5: Georaphical locations of all recorded scenes represented by their ego-poses shown in red.

attributes. While the distribution of the scene tags and the distribution of the bounding box categories can be seen in Figure 3 and Figure 4, respectively, Figure A6 shows the distribution of the object categories by dataset split. It can be seen that the different splits show a similar class distribution with slight variations due to the optimization criteria on temporal and spatial divergence.
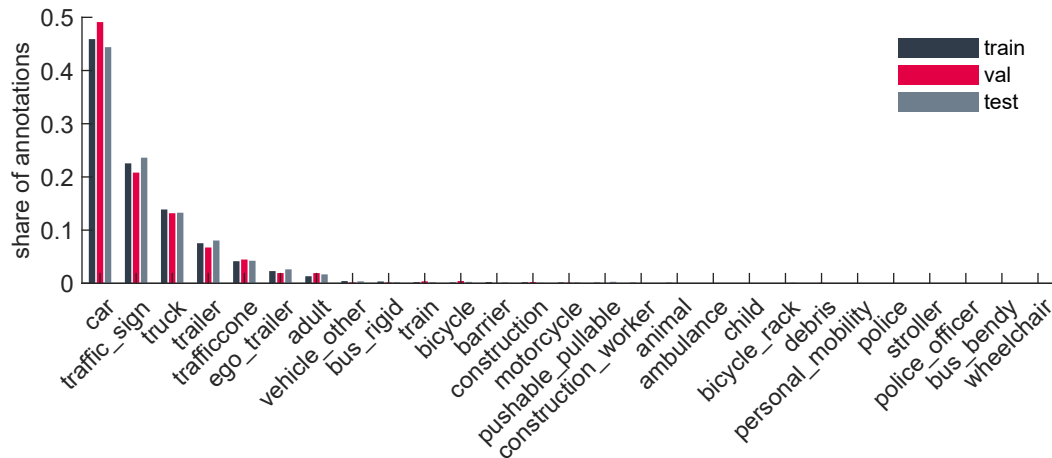
Figure A6: Distribution of the object categories grouped by dataset split.

In addition to the analysis of category and tag distribution, Figure A7 provides an analysis of the distribution of the different object attributes. It can be seen that most objects are moving and characterized by high visibility. Still, 39 % of the objects have a visibility of less than 41 % (visibility level 1). This number reflects the challenges of long-range perception and occlusions by large vehicles.

Another important measure for tracking and prediction is the tracking duration of different objects, which is why Figure A8 is included. The analysis shows that the mean tracking duration across all categories is 9.4 s, with a median duration of 8 s for the car class. Since established prediction metrics, like the nuScenes or Argoverse2 metric, use a prediction horizon of 6 s, the majority of all tracks are suitable for this purpose. It can also be seen that the ego-trailer class is reliably tracked throughout the whole scene (if a trailer is attached), which provides the possibility for developing trailer detection or articulation angle estimation methods.

In addition to the class and attribute distribution, we also analyzed the spatial distribution of the annotated objects. Figure A9 shows that the majority of the objects occur in the front or the back of the vehicle and that most of the objects are to the left of the ego vehicle. This is probably because a significant proportion of the scenes were recorded on highways, and (in Germany) vehicles must always use the rightmost lane whenever possible. Therefore, most of the other vehicles are either part of the oncoming traffic or overtaking our ego vehicle, which results in the shown distribution.

14

(a) Pedestrian

(b) Cycle

(c) Traffic Sign

(d) Vehicle

(e) Visibility
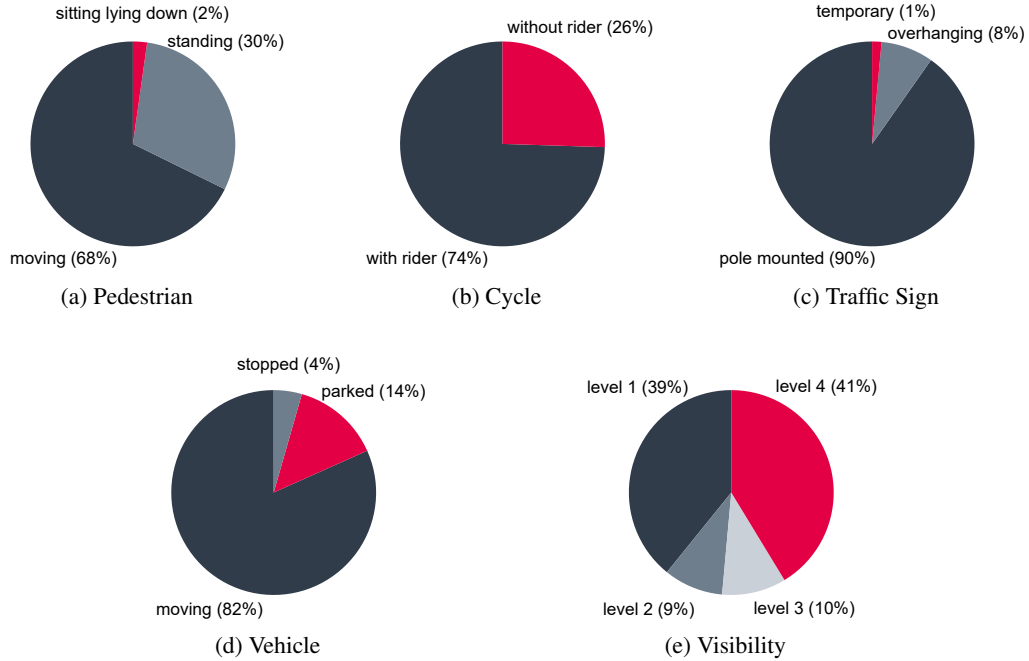
Figure A7: Distribution of the 15 different object attributes shown by the five attribute groups.
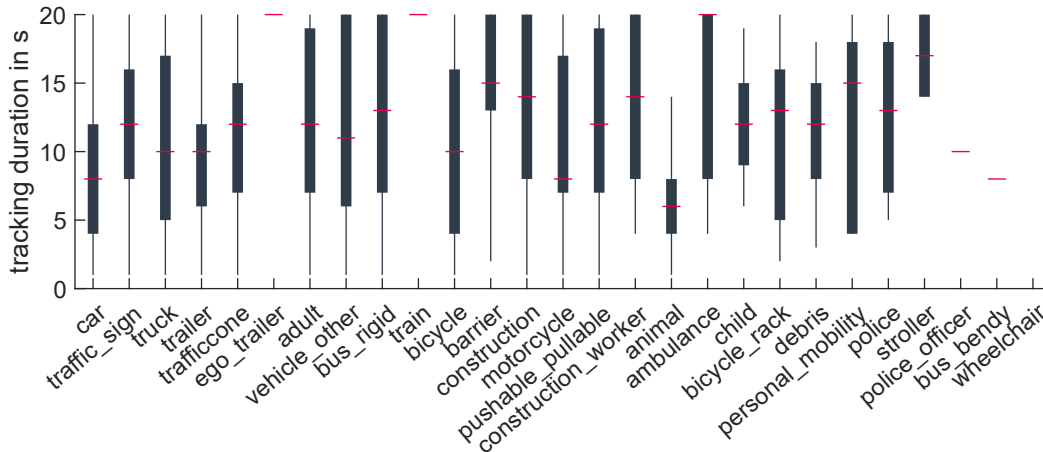
Figure A8: Instance tracking duration for all instances of the dataset represented by their median tracking duration (red lines), interquartile range (blue boxes), and whiskers (blue lines).

Another unique property of MAN TruckScenes is the provision of long-range detections and annotations of objects with high velocities. As listed in Table 1, the 99.9th percentile of all annotated bounding box distances is 226 m and 50 % of all annotations are beyond 75 m. This number is significantly higher than the 14 % of the Argoverse2 dataset, whereas nuScenes and Waymo have less than 1 % objects beyond 75 m [23]. A detailed distribution of the MAN TruckScenes annotation distances can be seen in Figure A10.

In addition, MAN TruckScenes includes a significant amount of objects with high absolute velocities. The average velocity of all objects within the dataset is 14 m/s and 40 % of all objects are moving faster than 20 m/s. The velocity distribution of all moving objects (objects with a velocity higher than 0.5 m/s) is shown in Figure A10. It can be seen that a significant amount of objects are moving faster than 30 m/s, while a local maximum can be observed at 24 m/s. This can be explained by the fact that, in Germany, the maximum permitted speed for vehicles heavier than 3.5 t is 80 km/h (22 m/s).

15
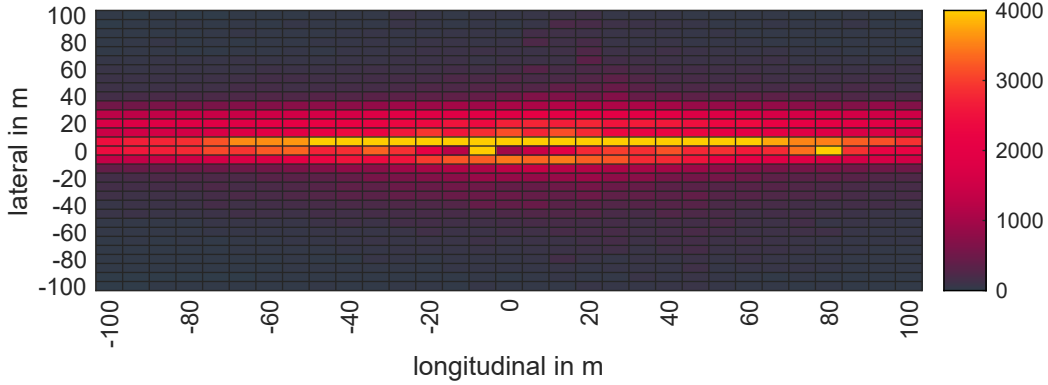
Figure A9: Spatial distribution of the annotated objects in the vehicle coordinate frame. The color indicates the number of annotated objects within the particular cell throughout the overall dataset.
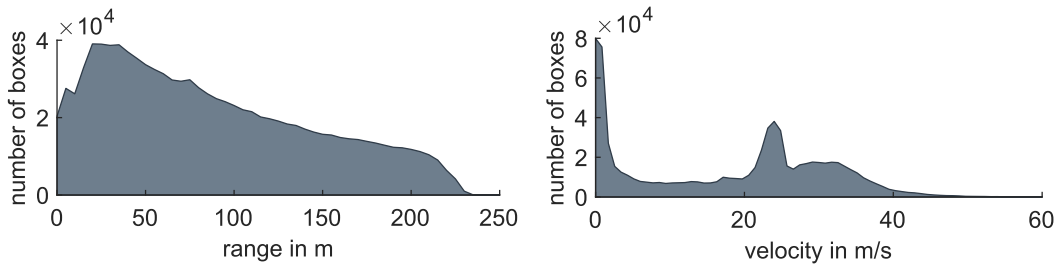
Figure A10: Number of annotated 3D bounding boxes by range (right) and by velocity (left). Note that the velocity distribution only considers moving objects with more than 0.5 m/s.

The MAN TruckScenes dataset provides an average of 34 objects per sample, which is comparable to the 33 objects per sample of the nuScenes dataset but less than the Argoverse2 or Waymo dataset with 75 and 61, respectively [23]. This is probably because the aforementioned datasets were recorded in city environments, while MAN TruckScenes was mainly recorded on highways and terminal environments. However, the validity of such a comparison is limited due to the different taxonomies of the datasets. Nevertheless, the sample-based object distributions of the MAN TruckScenes dataset are shown in Figure A11.
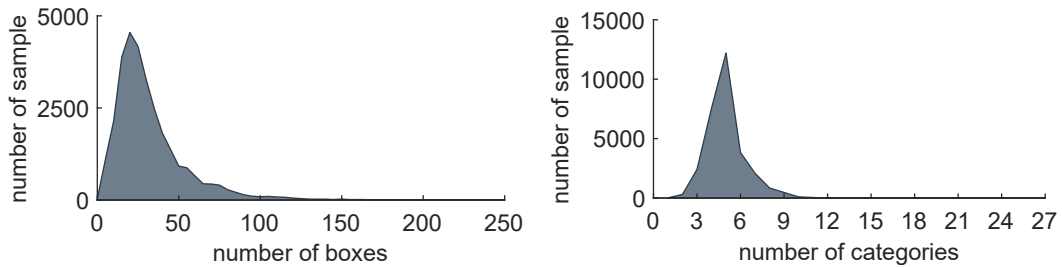
Figure A11: Number of dataset samples over the number of annotated 3D bounding boxes (right) and over the number of different object categories (left).

A.6 Experiments

Experiments were conducted for all three sensor modalities. The camera results are based on the PETR [16] architecture, the radar baseline on RadarGNN [10], and the lidar results on Center-Point [24]. To be specific, the camera baseline utilizes a PETR [16] model with a pre-trained FCOS3D (V-99-eSE) backbone with two multi-scale feature map outputs. Subsequently, a Feature Pyramid Network (FPN) neck is used for feature alignment and a transformer-based detection head is used with six decoder layers. The 900 query points are initialized within an area of $320\,\mathrm{m} \times 320\,\mathrm{m} \times 20\,\mathrm{m}$

16

and both the features as well as queries are positional encoded with a sine encoding. The training is Non Maximum Suppression (NMS) free and uses a Hungarian Assignment method instead, combined with a Focal Loss. The implementation is based on the MMDetection3D framework and uses common image augmentation techniques like random rotation, scaling, or translation. The final model is trained on an Nvidia A100 GPU for 48 epochs (38 h) and a batch size of 8.

The radar model uses the translation invariant RadarGNN [10] variant with a KNN-based graph construction method with a maximum of 20 neighbors in a radius of 1 m and directed edges with associated features. The main body of the model consists of five consecutive Message-Passing Neural Network (MPNN) layers and a separate classification and regression head. The training is based on a Cross-Entropy and Huber Loss function and does not utilize any data augmentation methods. The final training is conducted on an Nvidia A100 GPU with a batch size of 32 for 30 epochs.

The lidar method is based on a CenterPoint [24] model with a voxel-based encoder based on the SECOND architecture, a FPN neck, a Deep & Cross Network (DCN) CenterHead, and circular NMS. The training scheme uses different data augmentation methods, including random global rotation, translation, scaling, flipping, point shuffling, and object sampling. The overall model is based on the MMDetection3D framework and was trained for 20 epochs on two Nvidia A100 GPUs for 181 h with a batch size of 16 per GPU.

In the following, detailed results of the different models are provided for every sensor modality. Table A1 - A3 show the class-specific results per sensor modality, while Table A4 lists the object detection results over different range bins. Table A5 shows the detection results by scene tags and Figure A12 - A14 provides qualitative results under various environmental conditions.

Observations show that the camera model struggles to perform well on minority classes and performs generally better on larger than smaller objects. The results show lots of false positives (FP) and the accumulation of bounding boxes on drivable areas. Besides that, the camera model's performance decreases with increasing distance and performs significantly worse in the dark or tunnel environments.

The radar model, on the other hand, struggles to detect non-metallic objects but shows promising results on the majority classes. The results of the radar experiments also suggest that the accurate prediction of object orientations is most challenging and the velocity estimation seems to be off due to the utilization of the relative radial velocity instead of the ego-motion compensated velocity measurements. Furthermore, the radar-based detection performs significantly worse in the terminal environment and environments with a lot of radar reflections.

Lastly, the results of the CenterPoint model training show that the detection quality of diverse object classes (like 'animal', 'other vehicle', or 'bus') is insufficient, while more homogeneous object classes (like 'traffic sign', 'traffic cone', or 'pedestrian') have a higher detection score. Moreover, the size of the objects or the number of objects in the dataset are no sufficient predictors for the detection quality, as shown by the 'truck' class. In addition, the results suggest that the detection quality decreases with increasing distance and that rainy or foggy weather conditions negatively affect the model's performance. The model also performs significantly worse in tunnels or rural areas. However, safe autonomous trucking requires robust perception under all environmental conditions, which is why we release MAN TruckScenes to promote research in this area.

17

Table A1: Results of the camera 3D object detection task per detection class on the test split of the MAN TruckScenes dataset v1.0.

	PETR					
	AP	ATE	ASE	AOE	AVE	AAE
car	0.03	1.26	0.22	0.20	4.18	0.12
truck	0.01	1.27	0.35	0.11	2.17	0.20
bus	0.00	1.00	1.00	1.00	1.00	1.00
trailer	0.19	0.67	0.19	0.04	1.84	0.12
other_vehicle	0.00	1.00	1.00	1.00	1.00	1.00
pedestrian	0.00	1.25	0.48	1.19	0.30	0.63
motorcycle	0.00	1.00	1.00	1.00	1.00	1.00
bicycle	0.00	1.00	1.00	1.00	1.00	1.00
traffic_cone	0.03	1.31	0.51	-	-	-
barrier	0.01	1.46	0.90	0.07	-	-
animal	0.00	1.00	1.00	1.00	1.00	-
traffic_sign	0.00	1.29	0.61	0.50	-	0.02

Table A2: Results of the radar 3D object detection task per detection class on the test split of the MAN TruckScenes dataset v1.0.

	RadarGNN					
	AP	ATE	ASE	AOE	AVE	AAE
car	0.36	0.53	0.22	1.48	26.50	0.02
truck	0.20	0.87	0.42	1.61	20.67	0.02
bus	0.00	1.00	1.00	1.00	1.00	1.00
trailer	0.27	0.66	0.15	1.09	18.86	0.02
other_vehicle	0.00	1.00	1.00	1.00	1.00	1.00
pedestrian	0.00	1.00	1.00	1.00	1.00	1.00
motorcycle	0.00	1.00	1.00	1.00	1.00	1.00
bicycle	0.00	1.00	1.00	1.00	1.00	1.00
traffic_cone	0.00	1.00	1.00	-	-	-
barrier	0.00	1.00	1.00	1.00	-	-
animal	0.00	1.00	1.00	1.00	1.00	-
traffic_sign	0.02	0.64	0.92	1.27	-	0.08

Table A3: Results of the lidar 3D object detection task per detection class on the test split of the MAN TruckScenes dataset v1.0.

	CenterPoint					
	AP	ATE	ASE	AOE	AVE	AAE
car	0.33	0.29	0.15	0.07	3.95	0.17
truck	0.23	0.42	0.16	0.03	2.62	0.39
bus	0.06	0.49	0.11	0.18	4.28	0.01
trailer	0.34	0.42	0.15	0.03	3.74	0.37
other_vehicle	0.14	0.65	0.34	0.15	0.85	0.32
pedestrian	0.45	0.29	0.39	0.81	0.43	0.37
motorcycle	0.22	0.33	0.22	0.08	5.57	0.03
bicycle	0.06	0.27	0.27	0.46	2.13	0.13
traffic_cone	0.55	0.28	0.37	-	-	-
barrier	0.42	0.19	0.62	0.03	-	-
animal	0.00	1.00	1.00	1.00	1.00	-
traffic_sign	0.42	0.28	0.44	0.21	-	0.03

18

Table A4: 3D object detection results (mAP / NDS) by range on the MAN TruckScenes test set v1.0.

	Modality	0 m - 25 m	0 m - 50 m	0 m - 100 m	0 m - 150 m
PETR	camera	0.06 / 0.13	0.04 / 0.13	0.03 / 0.12	0.02 / 0.12
RadarGNN	radar	0.09 / 0.12	0.08 / 0.12	0.08 / 0.11	0.07 / 0.11
CenterPoint	lidar	0.48 / 0.52	0.40 / 0.48	0.30 / 0.43	0.27 / 0.41

Table A5: Results (mAP / NDS) of the 3D object detection task by scene tag on the test split of the MAN TruckScenes dataset v1.0.

		PETR camera	RadarGNN radar	CenterPoint lidar
	modality			
weather	clear	0.02 / 0.11	0.05 / 0.10	0.30 / 0.42
weather	overcast	0.02 / 0.11	0.09 / 0.12	0.19 / 0.36
weather	fog	0.05 / 0.13	0.15 / 0.15	0.15 / 0.20
weather	other_weather	- / -	- / -	- / -
weather	rain	0.03 / 0.11	0.09 / 0.12	0.16 / 0.23
weather	snow	- / -	- / -	- / -
weather	hail	- / -	- / -	- / -
area	terminal	0.02 / 0.09	0.00 / 0.04	0.30 / 0.32
area	parking	- / -	- / -	- / -
area	highway	0.02 / 0.11	0.09 / 0.12	0.19 / 0.31
area	city	0.02 / 0.07	0.02 / 0.06	0.18 / 0.29
area	residential	0.01 / 0.06	0.03 / 0.07	0.12 / 0.28
area	rural	0.03 / 0.10	0.07 / 0.11	0.06 / 0.15
area	other_area	- / -	- / -	- / -
daytime	morning	0.02 / 0.11	0.08 / 0.11	0.17 / 0.36
daytime	noon	0.02 / 0.12	0.06 / 0.11	0.33 / 0.44
daytime	evening	- / -	- / -	- / -
daytime	night	0.02 / 0.10	0.06 / 0.09	0.17 / 0.24
season	spring	- / -	- / -	- / -
season	summer	0.02 / 0.12	0.07 / 0.11	0.34 / 0.42
season	autumn	0.02 / 0.11	0.08 / 0.11	0.20 / 0.35
season	winter	0.03 / 0.10	0.05 / 0.10	0.12 / 0.26
lighting	illuminated	0.02 / 0.12	0.07 / 0.11	0.26 / 0.41
lighting	twilight	0.05 / 0.09	0.06 / 0.10	0.19 / 0.27
lighting	dark	0.01 / 0.09	0.02 / 0.09	0.18 / 0.24
lighting	glare	0.02 / 0.11	0.09 / 0.12	0.16 / 0.23
lighting	other_lighting	- / -	- / -	- / -
structure	regular	0.02 / 0.12	0.06 / 0.10	0.29 / 0.39
structure	underpass	0.04 / 0.12	0.11 / 0.13	0.13 / 0.28
structure	bridge	0.03 / 0.11	0.09 / 0.12	0.14 / 0.30
structure	overpass	0.03 / 0.12	0.09 / 0.12	0.24 / 0.31
structure	tunnel	0.03 / 0.08	0.07 / 0.09	0.12 / 0.16
const	unchanged	0.02 / 0.12	0.07 / 0.11	0.25 / 0.40
const	roadworks	0.04 / 0.12	0.10 / 0.11	0.23 / 0.28

19

(a) night   (b) snow   (c) terminal
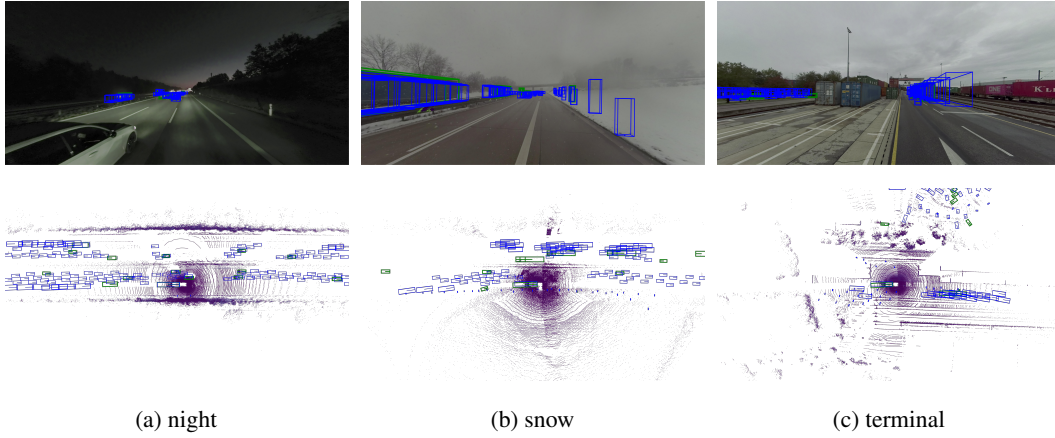
Figure A12: Example visualizations of the PETR model on the validation split of the MAN TruckScenes dataset v1.0. The front left camera is shown on the top and the fused lidar point cloud on the bottom. Model predictions are shown in blue and the ground truth in green.

(a) night   (b) snow   (c) terminal
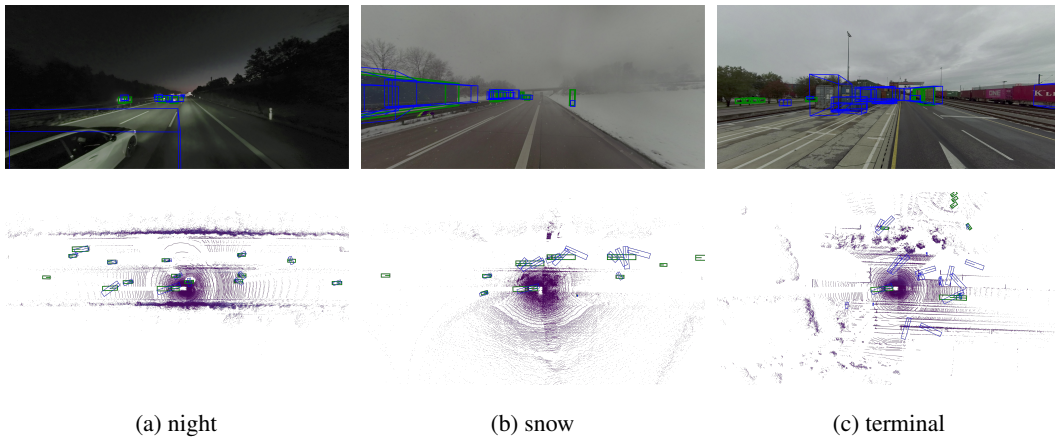
Figure A13: Example visualizations of the RadarGNN model on the validation split of the MAN TruckScenes dataset v1.0. The front left camera is shown on the top and the fused lidar point cloud on the bottom. Model predictions are shown in blue and the ground truth in green.

(a) night   (b) snow   (c) terminal
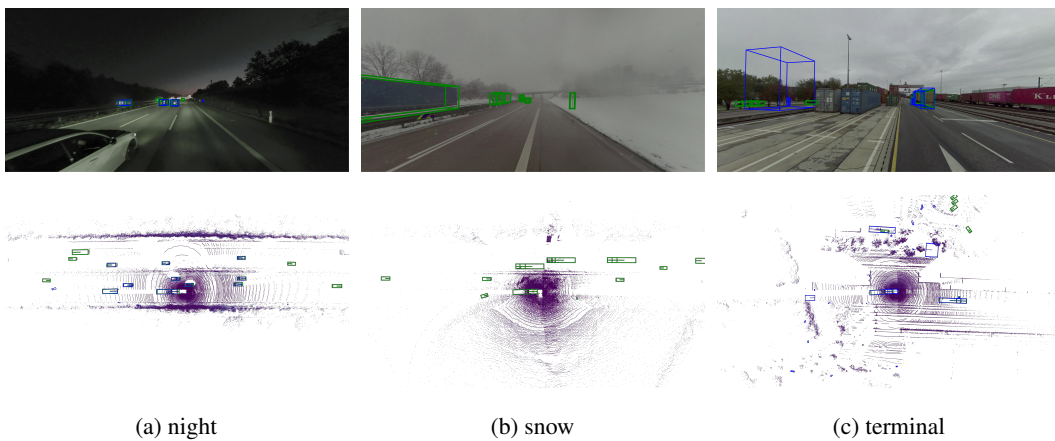
Figure A14: Example visualizations of the CenterPoint model on the validation split of the MAN TruckScenes dataset v1.0. The front left camera is shown on the top and the fused lidar point cloud on the bottom. Model predictions are shown in blue and the ground truth in green.

20

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [Yes] Limitations are addressed in Section 7.

    (c) Did you discuss any potential negative societal impacts of your work? [Yes] Dataset access and secure storage are part of the AWS sponsorship program. Biases and data distribution are discussed in Section 7 and societal impacts are discussed in Section 8.

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] The dataset is in line with European legislation and GDPR directives.

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [N/A]

    (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Data, code, and documentation are part of the supplemental material.

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Running the training multiple times is computationally too expensive.

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [Yes]

    (b) Did you mention the license of the assets? [Yes] The dataset is published under the CC BY-NC-SA 4.0 license and the code is published under the Apache 2.0 license.

    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Data, code, and documentation are part of the supplemental material.

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All data is recorded in accordance with European legislation and GDPR directives.

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The dataset is anonymized and in line with European GDPR directives, as described in Section 3.8.

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

21