Zenseact Open Dataset: A large-scale and diverse multimodal dataset for autonomous driving

Mina Alibeigi*†, William Ljungbergh*, Adam Tonderski*,
Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk,
Junsheng Fu, Jenny Widahl, Christoffer Petersson

Zenseact
first.last@zenseact.com

Abstract

Existing datasets for autonomous driving (AD) often lack diversity and long-range capabilities, focusing instead on 360° perception and temporal reasoning. To address this gap, we introduce Zenseact Open Dataset (ZOD), a large-scale and diverse multimodal dataset collected over two years in various European countries, covering an area 9× that of existing datasets. ZOD boasts the highest range and resolution sensors among comparable datasets, coupled with detailed keyframe annotations for 2D and 3D objects (up to 245m), road instance/semantic segmentation, traffic sign recognition, and road classification. We believe that this unique combination will facilitate breakthroughs in long-range perception and multi-task learning. The dataset is composed of Frames*,* Sequences*, and* Drives*, designed to encompass both data diversity and support for spatio-temporal learning, sensor fusion, localization, and mapping.* Frames *consist of 100k curated camera images with two seconds of other supporting sensor data, while the 1473* Sequences *and 29* Drives *include the entire sensor suite for 20 seconds and a few minutes, respectively. ZOD is the only large-scale AD dataset released under a permissive license, allowing for both research and commercial use. More information, and an extensive devkit, can be found at* `zod.zenseact.com`.

1. Introduction

Road traffic accidents cause more than 1.3 million deaths and many more nonfatal injuries and disabilities globally each year [20]. Automated driving has the potential to improve road safety by intervening in accident-prone situations or even controlling the entire ride from start to destination. Regardless of the level of automation, autonomous vehicles require sensors such as cameras, GNSS (Global Navigation Satellite System), IMU (Inertial Measurement Unit), and range sensors such as radar or LiDAR (Light

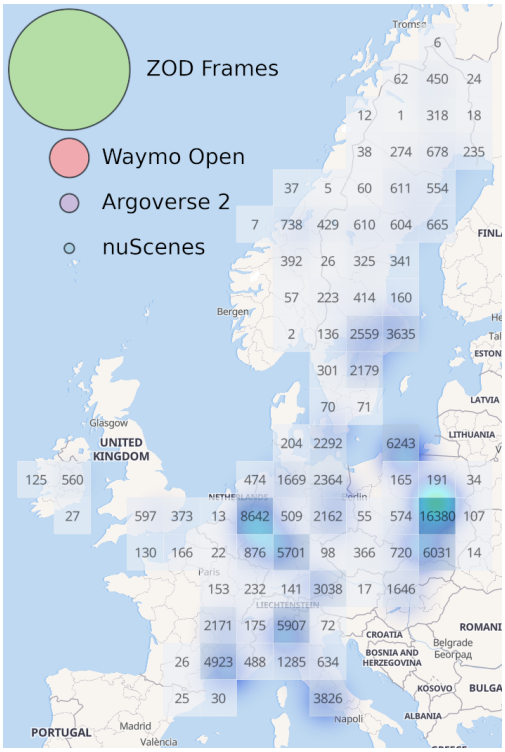


Figure 1: Geographical coverage comparison with other AD datasets using the diversity area metric defined in [25] (top left), and geographical distribution of ZOD *Frames* overlaid on the map. The numbers in the quantized regions represent the amount of annotated frames in that geographical region. ZOD *Frames* contain data collected over two years from 14 different European countries, from the north of Sweden down to Italy.

*Equal contribution.
†Corresponding author.

Detection and Ranging) to perceive their surroundings accurately. Moreover, they require advanced perception, fusion, and planning algorithms to make use of this data efficiently. Machine learning (ML) algorithms, particularly deep learning (DL), have been increasingly used to develop autonomous driving (AD) software, but they require high-quality and diverse data from real-world traffic scenarios to achieve the necessary performance.

The development of AD systems owes much of its recent success to the availability of large-scale image datasets [3, 8, 9, 16, 19, 34] and dedicated multimodal AD datasets [4, 6, 11, 12, 17, 25, 30, 32]. These AD datasets have emphasized temporal reasoning and 360° perception, as this corresponds to a nominal self-driving setup. However, as data from the same scene is highly correlated, this naturally limits diversity in terms of weather or lighting conditions, driving situations, and geographical distribution. This can result in overly specialized solutions, which may not generalize to the full operational design domain of real-world AD systems.

To complement existing datasets, we introduce Zenseact Open Dataset (ZOD), Europe's largest and most diverse multimodal AD dataset. ZOD consists of more than 100k traffic scenes that have been carefully curated to cover a wide range of real-world driving scenarios. The dataset is split into three subsets: 100k independent *Frames*, 1473 twenty-second *Sequences*, and 29 *Drives* lasting a few minutes. *Frames* are primarily suitable for non-temporal perception tasks, *Sequences* are intended for spatio-temporal learning and prediction, and *Drives* are aimed at longer-term tasks such as localization, mapping, and planning. This separation allows ZOD to cover an area $9\times$ larger than any other AD dataset, offering ample opportunities for developing robust algorithms that generalize well across multiple operational domains. We also facilitate research in domain adaptation and transfer learning by providing comprehensive metadata for each scene.

Robust performance in various conditions, including high speeds, will be vital for the deployment of AD systems. In particular, high-speed scenarios puts a hard requirement on long-range perception, which in turn puts challenging requirements on sensor resolution. ZOD stands out from other AD datasets by employing high-resolution sensors, such as an 8MP camera, rooftop LiDARs with 254k points per scan, and a high-precision GNSS/IMU inertial navigation system with 0.01m position accuracy. Additionally, we provide manual keyframe annotations for several perception tasks, such as semantic and instance segmentation masks for roads and lanes, 2D and 3D bounding boxes for static and dynamic objects (up to 245 meters), and road condition labels. We further annotate traffic signs with a rich taxonomy of 156 classes. The 446k unique labeled instances constitute the largest traffic sign dataset to date. We believe that the combination of high-quality sensors and detailed annotations in ZOD will enable breakthroughs in accurate and long-range perception, which is crucial for high-speed driving scenarios.

ZOD's extensive annotations for multiple perception tasks make it an ideal dataset for multi-task learning, which is a recent trend in computer vision and AD [26, 35, 36]. The core idea of multi-task learning is to learn a shared representation that can benefit all tasks, resulting in improved generalization and performance on individual tasks [5]. This approach also allows models to make better use of data and available resources, a crucial feature for real-world applications such as AD systems as they typically operate on embedded hardware with limited computational power.

To ensure the privacy of individuals and comply with legal and regulatory requirements, we employ two approaches to anonymize faces and license plates: blurring and replacement with synthetic data. These anonymization techniques were chosen to enable research on the impact of anonymization techniques on learning methods, with initial results demonstrating that none of the techniques have a negative impact on performance.

Finally, ZOD is the first large-scale AD dataset released under the permissive CC BY-SA 4.0 license [1]. This license allows for research and commercial use (subject to the license terms), as well as sharing and adapting permits, which provides an opportunity for startups and other commercial entities to leverage the dataset for their projects. We believe that this open and inclusive approach will foster innovation and accelerate the development of AD technology beyond the research community. To facilitate a rapid start with ZOD, it comes with an extensive development kit, including multiple tutorials and examples.

In summary, our main contributions are the following:

- We release ZOD, the most diverse autonomous driving dataset to date. The data is collected from Europe over multiple years and is curated to contain a wide range of traffic scenarios, weather conditions, road types, and lighting conditions.
- The data is collected using high-resolution sensors and coupled with detailed keyframe annotations for 2D/3D objects, lane instances, and road segmentation, enabling long-range perception with annotated objects farther away than in any other comparable AD dataset.
- The object annotations include a rich traffic sign taxonomy with more than twice as many unique instances as the largest existing traffic sign dataset.
- ZOD is the first large-scale AD dataset released under a permissive license, allowing both research and commercial use.

Dataset	Locations	Geo. coverage	Ann. frames	Sequences	Size (hr)	Ann. range†	Avg. LiDAR points	Camera	Map
KITTI [11]	Karlsruhe	-	15k⋆	22	1.5	91 m	120k	90°	No
nuScenes [4]	Boston, Singapore	5 km²§	40k⋆	1000	5.5	141 m	34k	360°	**Yes**
ONCE [17]	China	-	16k⋆	581	**27.8**	81 m	65k	360°	No
PandaSet [32]	San Francisco	-	8k⋆	103	0.2	**300 m**	166k	360°	No
Waymo Open [25]	6 U.S. cities	76 km²§	**400k**⋆	**2030**	11.3	80 m	177k	360°	No
A2D2 [12]	3 German cities	-	12k	-	-	103 m	7k	360°	No
Argoverse 2 [30]	6 U.S cities	17 km²	150k⋆	1000	4.2	214 m	107k	360°	**Yes**
ZOD *Frames*	**14 European countries**	**705 km²**	100k	-	**55.6**‡	245 m	**254k**	120°	No
ZOD *Sequences*	6 European countries	26 km²	1473	1473	8.2	245 m	**254k**	120°	No
ZOD *Drives*	2 European countries	-	-	29	1.5	-	**254k**	120°	No

Table 1: Dataset comparison. Geographical coverage is computed over annotated frames, § refers to values taken from [25]. ⋆Sequential annotation, allowing temporal tasks. ‡Counting each *Frame* as a 2-second LiDAR sequence, note that only the center image is provided. †Showing the 99.9th percentile computed using the publicly available data.

2. Related work

Over the last decade, the development of AD datasets has been a focus area to advance AD research and enhance road safety. Many of these datasets have been devoted to visual perception [3, 8, 19, 34]; However, in recent times, multi-modal datasets have become increasingly popular as most AD systems aim to fuse information from various onboard sensors to generate a robust representation of the world and make more informed decisions.

KITTI [11] is widely regarded as one of the most impactful multimodal AD datasets to date. Released in 2012, KITTI provides 22 road sequences with stereo cameras, LiDAR, and high-precision GNSS/IMU sensor data from real-world driving in Karlsruhe, Germany. With 200k object labels in the form of 3D tracklets, KITTI enabled significant advancements in AD research, including 3D object detection and tracking, visual odometry, and scene flow estimation. However, as modern algorithms and AD systems tackle more complex tasks, there is a growing demand for larger, more diverse, and more advanced datasets.

Since the release of KITTI, the field of AD research has seen several large-scale AD datasets striving to push the boundaries in various aspects. Among these is the ApolloScape dataset [29], which boasts one of the largest publicly available collections of annotated video frames for semantic segmentation in AD, with over 140k frames. However, the dataset's strong temporal connection between frames limits the diversity of training data. The KAIST multispectral dataset [7] draws attention for its use of thermal cameras, while H3D [21] stands out as one of the first datasets to fully annotate 3D objects in a 360° perspective. A2D2 [12] is one of the first public AD datasets allowing commercial use but with restrictions on sharing the derivates. It offers RGB images, LiDAR data, and vehicle bus data, with 41k frames containing semantic segmentation labels and 12k frames with 3D bounding box labels for objects visible in the front camera's field of view. However, A2D2 suffers from extremely sparse point clouds. PandaSet [32] is also released under a permissive license, and thanks to their multi-LiDAR setup, it offers much denser point clouds which allow them to annotate 3D bounding boxes up to 300m. Similar to A2D2, they also offer RGB images and vehicle bus data, together with annotations across their 8k frames. While being multi-modal datasets, both A2D2 and PandaSet suffer from low annotation counts, and similarly to H3D and KAIST, their limited size has hindered widespread adoption.

NuScenes [4], released in 2019, became a significant milestone in AD research as one of the first publicly available large-scale datasets. It provides a rich surround view by means of comprehensive sensor data from LiDAR, six cameras, and five radars. Additionally, the dataset includes semantic maps of roads and sidewalks, vehicle bus data, and GNSS information. NuScenes is organized into sequences of 20 seconds and provides detailed temporal annotations for 3D objects at 2Hz, making it one of the first datasets to do so. This feature has enabled many breakthroughs in the development of algorithms for detection, tracking, prediction, and planning.

Waymo Open Dataset [25] and Argoverse 2 [30] are two large-scale datasets that adopt the strengths of nuScenes, while addressing many of its weaknesses. Both datasets cover a more comprehensive range of driving scenarios and include higher-resolution sensors. Waymo's dataset excels in geographical diversity, while Argoverse 2 stands out for its long-range annotations, maps, and extensive object taxonomy. While these three datasets push the state-of-the-art in AD, especially in terms of spatio-temporal reasoning and surround vision, they do so at the cost of high sample correlation.

In contrast, ZOD makes a different set of trade-offs. Our *Frames* subset has nine times larger geographical coverage than the Waymo Open Dataset, with only a quarter of the frames. By complementing existing datasets, our goal is to facilitate breakthroughs in long-range perception and multi-task learning and enable the benchmarking of algorithms for both research and commercial use.

Sensors	Details
LiDARs	1xVelodyne VLS128 rotating 3D laser scanner, HFOV* 360°, VFOV* 40° [-25°, +15°], HRES* 0.1° to 0.4°, VRES* 0.11°, channels 128, wavelength 903 nm, range up to 245 m, and frame rate 10 Hz. 2xVelodyne VLP16, HFOV 360°, VFOV 30° [-15°, +15°], HRES 0.1° to 0.4°, VRES 2°, channels 16, wavelength 905 nm, range up to 100 m, frame rate 10 Hz, collecting up to 0.3 million points/second.
Camera	1xRGB front-looking camera, HFOV 120°, VFOV 67°, and resolution 3848x2168 (8MP).
High precision GNSS/IMU	1xOxTS RT3000 inertial and GNSS navigation system, six axes, L1/L2 RTK with frame rate 100 Hz, 0.01m position accuracy, 0.03° pitch/roll and 0.1° heading accuracy.

* HFOV/VFOV and HRES/VRES represent the horizontal/vertical field of view and resolutions, respectively.

Table 2: Sensor specifications of ZOD.

Another recent trend in the AD community is to release large-scale datasets without annotations, primarily for self-supervised learning purposes. The ONCE dataset [17] is one such example, containing a million LiDAR scenes and their corresponding camera images collected over three months in various areas, lighting, and weather conditions throughout China. However, the LiDAR point clouds in ONCE are limited to 65k points per frame, and the dataset lacks localization information, such as GNSS or map data. Similarly, the Argoverse 2 Lidar [30] dataset contains a vast number of unlabeled LiDAR sequences; however, it does not include camera data, which is essential for many AD tasks. To this end, ZOD includes *Sequences* and *Drives* subsets that are mostly unlabeled and consist of diverse traffic scenes with the camera, LiDAR, high-precision, and consumer-grade GNSS/IMU data.

Traffic sign recognition is another essential research avenue for enabling AD systems, that is typically pursued separately. The German Traffic Sign Benchmark Dataset [24] is one of the first datasets created to facilitate research on the classification of traffic signs. Multiple regional datasets [14, 18, 23, 37] have followed since, containing a varying number of images, signs, and classes. The Mapillary Traffic Sign Dataset (MTSD) [10] is one of the largest and most diverse traffic sign datasets to date, containing 206k labeled sign instances and a taxonomy of 313 classes from all over the world. Although ZOD has a smaller taxonomy of 156 classes, it has twice the amount of labeled sign instances (*i.e.*, 446k) compared to MTSD.

We refer the reader to Table 1 for a comprehensive comparison of ZOD with other multimodal AD datasets.

3. Zenseact Open Dataset

ZOD is a multimodal dataset containing a variety of real-world traffic scenes from highways, urban areas, and country roads around Europe. The dataset is collected under diverse weather conditions (clear, cloudy, rainy, and snowy) and lighting conditions (day, night, and twilight) over two years. It contains an extensive and detailed collection of fine-grained annotations for various tasks, including semantic and instance segmentation for the road, 2D and 3D bounding boxes for the dynamic and static objects including traffic signs with a rich taxonomy, and road classification labels.

The subsequent sections provide an elaborate description of the sensor suite, data, and annotations, followed by a statistical analysis of the dataset in Section 4.

3.1. Sensor suite

The data collection has been conducted using several vehicles with an identical sensor layout driven around Europe over the course of two years. The cars are equipped with a high-resolution camera, front- and side-looking LiDARs, and high-precision GNSS/IMU sensors. Figure 2 illustrates the sensor locations and their respective coordinate systems, whereas Table 2 gives comprehensive sensor specifications.
LiDAR data: The LiDAR point clouds are captured at 10 Hz and stored in a standard binary file format (`.npy`) per scan. Each file contains data from all three LiDAR sensors (VLS128 and VLP16s), represented as a 6-dimensional vector with the timestamp, 3D coordinates (x, y, and z), intensity, and diode index. The timestamp is relative to the frame timestamp in UTC, and the 3D coordinates are in me-

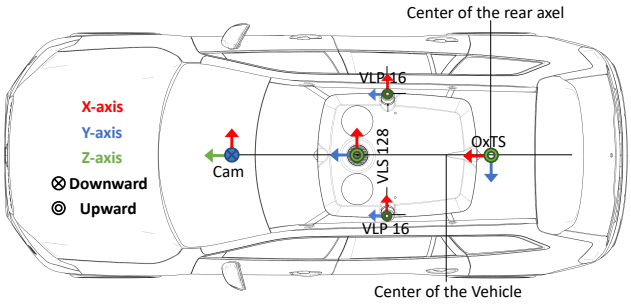


Figure 2: Placement of sensors used by data collection vehicles in ZOD and their corresponding coordinate systems.

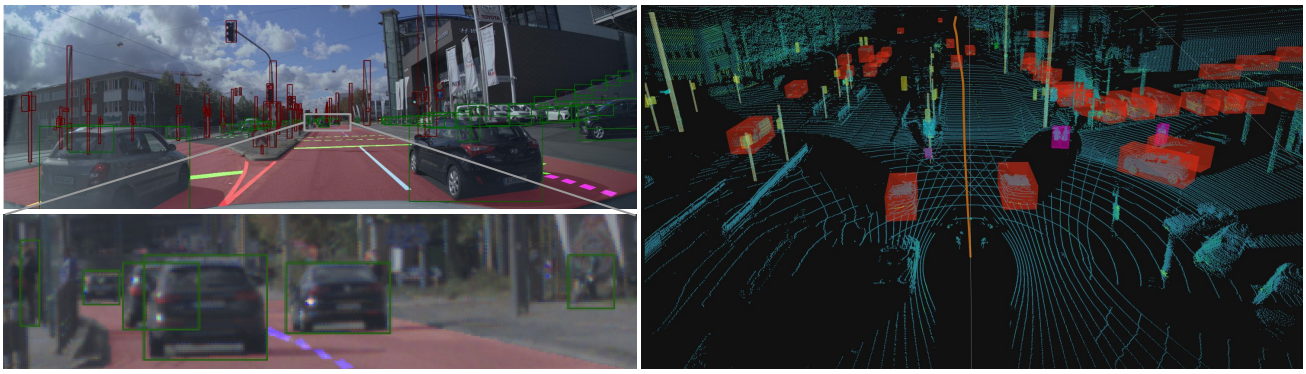


Figure 3: One *Frame* overlaid with multi-task annotations (top-left) and a zoomed-in version (bottom-left), highlighting annotated dynamic objects and lane instances at long distances. We also show the LiDAR point cloud with annotated 3D boxes and 200 m of future trajectory for the ego-vehicle (right) from the same *Frame*.

ters. Intensity is a measure of the reflection magnitude ranging from 0 to 255, and the diode index specifies the emitter that produced the point. Each LiDAR point cloud contains around 254k points on average and can be easily read and visualized using the provided development kit.

High-precision GNSS/IMU data: The high-precision GNSS/IMU data is logged at 100Hz and stored as HDF5 files, including UTC timestamp in seconds, WGS84 geographic coordinates (latitude, longitude, and altitude), ECEF Cartesian coordinates, heading, pitch, roll, velocities, accelerations, angular rates, and satellite information. This data can be used as ground truth for training different ML models, such as ego-vehicle trajectory prediction. A comprehensive description of the fields, coordinate transformation, and visualization functionalities (like the ego-vehicle trajectory in Figure 3) can be found in the development kit.

Camera data: The camera data is captured by high-resolution (8MP) wide-angle fish-eye lenses. All raw captured camera data is converted to RGB images using an internal production-level image signal processor. The RGB camera images are captured at 10Hz and saved as JPG files for a more accessible download of ZOD. However, we also provide lossless-PNG camera images. Considering the marginal compression impact on the learning models (see the supplementary material for empirical evidence), we strongly suggest using JPG images for benchmark experiments on ZOD.

Vehicle data: Various vehicle data are also released for *Sequences* and *Drives*. These include vehicle control signals such as steering wheel angle, acceleration/brake pedal ratios, and turn indicator status, as well as consumer-grade IMU and satellite positioning data. The vehicle control signals, IMU, and satellite positioning data are logged at 100Hz, 50Hz, and 1Hz, respectively.

3.2. Calibration and coordinate systems

To avoid drift over time and to achieve good cross-modality data alignment, all sensors are carefully and regularly synchronized and calibrated with regard to the specified ISO-8855 reference coordinate system during the data collection process. The origin of the reference coordinate system is at a fixed point relative to the vehicle chassis such that it appears in the center of the rear axle, given identical load conditions as during calibration. Under these conditions, it has the axes X-forward, Y-left, and Z-up. Individual sensor calibrations are provided for each datapoint, containing all the required information (*e.g.*, intrinsic and extrinsic calibration) to transform data between any two sensors. Moreover, functionalities for coordinate transformations and projections are available in the development kit.

3.3. Privacy protection

To protect the privacy of every individual in our dataset, and to comply with privacy regulations such as the European Union's General Data Protection Regulation (GDPR) [27], we use third-party services [2] and anonymize all objects in the images that contain personally identifiable information, *i.e.*, human faces and vehicle license plates.

Two anonymized RGB images are provided per *Frame* in ZOD, one using blurring and one using Deep Neural Anonymization (DNAT), see Figure 3 for examples of license plate anonymization using DNAT. The latter is based on generative AI, has minimal pixel impact, and maintains information like the line of sight of pedestrians while preserving anonymity. To the best of the authors' knowledge, ZOD is the only dataset that provides two anonymized versions of the original camera images, enabling an impact study of anonymization approaches on the quality of ML

models (Section 4.4) while supporting use cases such as human intent prediction in a compliant way.

3.4. Data categories

ZOD is categorized into three groups: *Frames*, *Sequences*, and *Drives*. The following subsections describe the content of each category, and Section 3.1 gives details per sensor data.

Frames: We carefully curate and select 100k frames from all over Europe, representing diverse traffic, location, weather, and lighting conditions. Each *Frame* scene contains two anonymized versions of an RGB camera image (*i.e.*, blurred and DNAT), captured from a front-looking camera mounted at the top of the windshield. We define the camera images as *keyframes*, considering they are fully annotated for different tasks (Section 3.5). Moreover, a ± 1-second sequence of LiDAR scans at 10Hz frequency is provided around each *keyframe*. This has been complemented by adding high-precision GNSS/IMU data at a frequency of 100Hz, which covers five seconds before and either 25 seconds after the *keyframe* or 300m ahead, whichever occurs first. To facilitate the extraction of interesting custom scenarios, a list of metadata describing the timestamp, geographical position, country code, weather conditions (*e.g.*, clear, rainy, foggy, snowy), solar elevation angle, road type (*e.g.*, highway, city), and the total number of annotated objects (*e.g.*, vehicles, pedestrians, vulnerable vehicles) is also provided for each *Frame*.

Sequences: We select an additional 1473 varied scenes, named *Sequences*, each with a duration of 20 seconds. *Sequences* can be utilized in applications that require temporal reasoning, such as visual odometry and ego-vehicle trajectory prediction. After the anonymization impact study on *Frames* (Section 4.4), we release only the blurred anonymized version of RGB camera images recorded at 10Hz for each sequence. Similar to *Frames*, LiDAR scans and high-precision GNSS/IMU data are also provided per scene, but for the entire 20 seconds. We also offer vehicle data for each *Sequence* scene in ZOD. This could be of particular interest to robotics research and higher-level scene understanding applications. Furthermore, the middle frame in each *Sequence* scene is carefully annotated for different tasks (Section 3.5), enabling spatio-temporal learning and automatic annotation generation tasks, among others.

Drives: We provide 29 *Drive* scenes, spanning a few minutes each, from two different cities and contain the same sensor data as the *Sequence* category. *Drive* scenes are deliberately chosen to capture changing road structures (*e.g.*, straight, curvy, banked, T-crossings, roundabouts, splits, merges, on-ramps, off-ramps), different road types (*e.g.*, urban, suburban, highways), and various traffic scenarios (*e.g.*, lane changes, cut-ins) to represent real-world complexities. They also encompass a couple of loop closures.

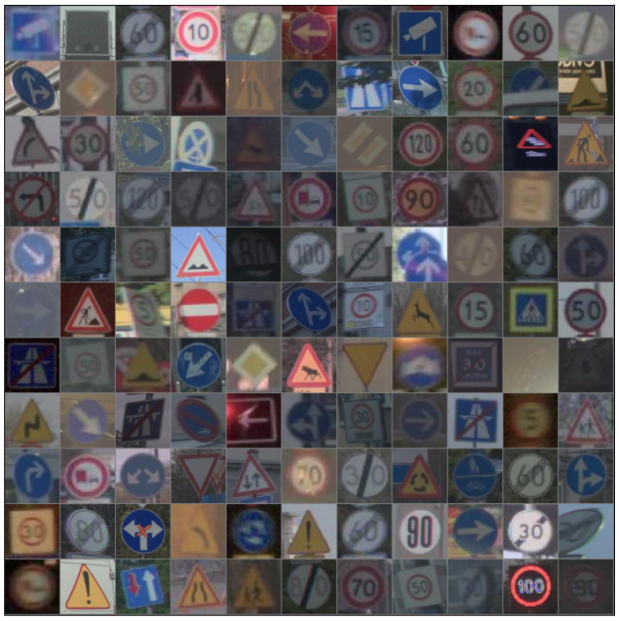


Figure 4: A random selection of cropped traffic signs, extracted from *Frames*. ZOD contains many difficult cases caused by occlusion, distance, viewing angle, lighting, *etc*.

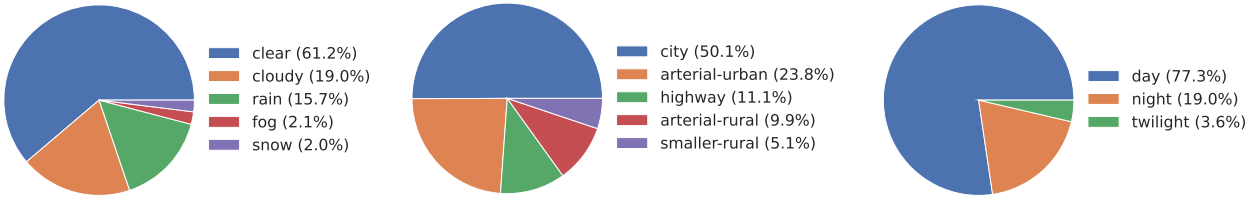
The *Drives* aim to be useful for research areas such as visual-based localization or simultaneous localization and mapping.

3.5. Annotations

We provide large-scale, high-quality, and detailed ground truth labels for several AD tasks per each *keyframe* in the *Frame* and *Sequence* scenes. All labels are manually created by skilled human annotators using commercial labeling tools and further passed through quality checks. The annotations are separated into three main categories: 1) semantic/instance segmentation masks for lane markings, road paintings, and ego road (a.k.a. driveable area), 2) 2D/3D bounding box labels for dynamic and static objects including traffic signs, and 3) classification labels for the road surface.

The first category of annotations consists of pixel-wise semantic segmentation labels with 15 top-level classes, see the supplementary material for the exact taxonomy. Instance segmentation labels are also provided for the annotated lane markings, along with additional properties such as color and cardinality, *i.e.*, if a lane marking is single or part of a group of lane markings. Figure 3 illustrate examples of lane marking and ego road annotations overlaid on the camera image.

All static and dynamic objects present in the camera images are annotated with a tightly fitting 2D bounding box indicated by the pixel coordinates of its four outermost points.



(a) Weather (b) Road type (c) Time of day

Figure 5: Distribution of weather (a), road types (b), and time of day (c) in ZOD *Frames*.

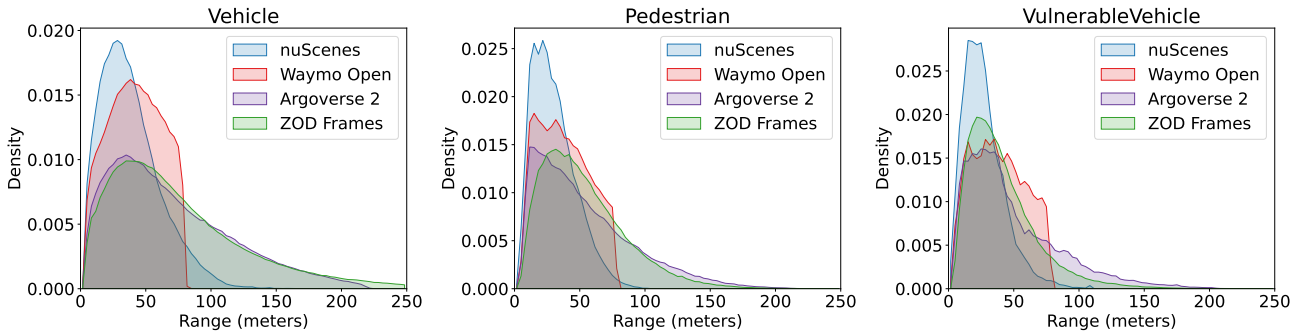


Figure 6: Distribution of the distance to the ego-vehicle over all annotated objects in nuScenes, Waymo Open, Argoverse 2, and ZOD *Frames*. The distribution of ZOD *Frames* is on-par with Argoverse 2. However, the annotated objects in ZOD *Frames* are less temporally correlated.

Objects visible both in the camera image and the LiDAR point cloud are also labeled with a 9-DOF 3D bounding box, described by the coordinate of the center of the box, length, width, height size, and the four quaternion rotation parameters of the cuboid (*i.e.*, qw, qx, qy, and qz). ZOD uses a hierarchical taxonomy, where dynamic objects are classified into four higher-level classes, which are further broken down into 16 subclasses. Static objects are classified into seven higher-level classes, which are broken down into 13 subclasses. Traffic signs are treated specially and are further labeled using an extensive taxonomy with 6 larger categories and 156 granular classes (Figure 4). By using this hierarchical taxonomy, ZOD provides a detailed annotation schema that enables researchers to analyze the dataset at different levels of granularity. Furthermore, objects are assigned properties, some generic (*e.g.*, occlusion rate), and some class-specific (*e.g.*, is_electronic for traffic signs or emergency for vehicles). Figure 3 shows examples of the annotated dynamic and static objects.

Lastly, we also provide details of the ego road surface condition (*i.e.*, wet or covered with snow) as road classification labels. Exact annotation taxonomies are detailed in the supplementary material.

4. Dataset analysis

In the following subsections, we analyze ZOD and highlight some key characteristics, namely its diversity (Section 4.1) and long-range objects (Section 4.2), and compare these with existing datasets. Moreover, we show the long-tailed nature of our dataset (Section 4.3) and, lastly, we analyze the impact that different anonymization techniques have on downstream computer vision tasks (Section 4.4).

4.1. Diversity

Most AD datasets are collected from a handful of cities, see Table 1. From the same table, it is also evident that most of Europe is underrepresented. To address this lack of diversity, we carefully curate data across Europe, ranging from the snowy parts of northern Sweden to the sunny countryside of Italy. In total, we provide data across 14 different countries. To quantitatively evaluate the geographical diversity of our dataset, we use the diversity area metric defined in [25] as the union of all 75m (radius) diluted ego-poses in the dataset. Using this definition, ZOD *Frames* obtains an area metric of 705km^2, which could be compared to 5km^2, 17km^2, and 76km^2 for nuScenes [4], Argoverse 2 [30], and Waymo Open [25], respectively[1]. Furthermore, we argue

[1]The geographical coverage values for the nuScenes and Waymo Open datasets are taken from [25].

Pipeline	Mode	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l	AP_{veh}	AP_{VV}	AP_{ped}
Faster-RCNN	original	30.23 ± 0.09	54.79 ± 0.06	28.72 ± 0.15	7.23 ± 0.04	30.49 ± 0.14	51.23 ± 0.07	42.41 ± 0.07	25.96 ± 0.15	22.32 ± 0.04
	DNAT	30.28 ± 0.03	54.86 ± 0.09	28.82 ± 0.08	7.15 ± 0.13	30.57 ± 0.06	51.31 ± 0.08	42.48 ± 0.02	25.90 ± 0.14	22.44 ± 0.03
	blur	30.17 ± 0.06	54.66 ± 0.10	28.83 ± 0.08	7.24 ± 0.02	30.50 ± 0.10	51.08 ± 0.10	42.41 ± 0.04	25.87 ± 0.13	22.22 ± 0.04
YOLOv7	original	33.62 ± 0.04	62.85 ± 0.03	30.84 ± 0.11	13.39 ± 0.09	35.81 ± 0.03	47.02 ± 0.21	47.79 ± 0.02	25.51 ± 0.04	27.56 ± 0.10
	DNAT	33.74 ± 0.09	62.95 ± 0.03	31.10 ± 0.11	13.54 ± 0.08	35.92 ± 0.14	47.17 ± 0.17	47.85 ± 0.11	25.67 ± 0.11	27.71 ± 0.08
	blur	33.67 ± 0.01	62.91 ± 0.07	30.90 ± 0.01	13.51 ± 0.06	35.92 ± 0.03	47.03 ± 0.02	47.89 ± 0.04	25.58 ± 0.12	27.53 ± 0.05

Table 3: Impact of image anonymization. We report AP (computed according to the COCO evaluation protocol [16]) when training image-based object detectors on images anonymized using three separate methods: None (original), DNAT, and blurring, while evaluation is done using the original images. The results show the mean and standard deviation across three separate runs. AP_{veh}, AP_{VV}, and AP_{ped} refer to AP for the vehicle, vulnerable vehicle, and pedestrian classes, respectively.

that – as our sensors and annotations range way beyond 75m – we can alter the definition to include the union of all 150m radius circles around the ego-poses, in which case ZOD *Frames* spans an area of 2039km^2 which is 26$\times$ the area covered in [25]. Importantly, we only include annotated frames when computing these metrics. We also show the entire geographical distribution of ZOD *Frames* in Figure 1.

To further illustrate the diversity of ZOD, we analyze the distribution of *Frames* across different weather conditions, road types, and times of day. As shown in Figure 5, our dataset includes frames captured in various weather conditions, including clear, cloudy, rainy, foggy, and snowy conditions. Moreover, our dataset covers different road types, including highways, urban roads, rural roads, and city streets, enabling the development and evaluation of models that can operate in different driving scenarios. Finally, we note that our dataset includes frames captured during different times of day, including almost 20k nighttime scenes, providing a comprehensive representation of real-world driving scenarios. By considering these diverse attributes during curation, we ensure that our dataset covers a broad range of driving conditions, enabling the development and evaluation of robust AD systems.

4.2. Long-range perception

As explained in Section 4.1, ZOD is a highly diverse dataset containing data from various driving conditions, ranging from slow-moving city driving to high-speed highway driving. To operate a vehicle safely when driving at speeds up to 130km/h (maximum ego-vehicle speed in ZOD is 133km/h), it is crucial to detect objects not only in your vicinity, but also at longer distances. To accurately perceive the environment at distances required for high-speed driving, the ego-vehicle has to be equipped with sensors with sufficient resolution to enable long-range perception. In ZOD, we have an 8MP front-looking camera coupled with high-resolution LiDAR sensors, allowing annotation of objects up to 245 meters away. This is – to the best of the authors' knowledge – farther than any other publicly available AD dataset of comparable size. In Figure 6, we show the distribution of the distance to the annotated 3D objects for three top-level classes, namely vehicles, vulnerable vehicles, and pedestrians. Note that the classes of the other datasets have been mapped to these three top-level classes. In particular, ZOD exhibits a range distribution similar to the Argoverse 2 dataset, with the exception of having a longer tail for vehicles and a lower density for distant vulnerable vehicles.

Table 4: 3D object detection performance by range.

	0-150m	0-50m	50-100m	100-150m	150-250m
mAP	0.25	0.33	0.19	0.06	0.01
CDS	0.46	0.62	0.33	0.08	0.02

Table 5: Traffic sign classification metrics [%].

F_1^{macro}	F_1^{micro}	$Acc_{avg}^{\uparrow 10}$	$Acc_{avg}^{\downarrow 10}$
78.5	95.4	93.4	65.4

We also aim to characterize the difficulty of 3D object detection on *Frames* by employing the widely used CenterPoint [33]. We trained CenterPoint and evaluated its performance across various range bins using the mAP and CDS metrics, as outlined in [30][2]. A comprehensive breakdown of the results is presented in Table 4. Our findings indicate that the detector struggles significantly at extended ranges, emphasizing the need for improved detectors and/or more relevant metrics. These challenges can be addressed using ZOD.

4.3. Long-tail perception

A critical dimension of autonomous driving is managing infrequent and challenging situations, including the precise detection of uncommon objects like wheelchairs and strollers. In examining this, we evaluate CenterPoint's performance across an extensive array of classes over the full 250m distance, as illustrated in Figure 7. While cars are

[2]The 0-150m bin corresponds to Argoverse 2's evaluation range, where CenterPoint achieves mAP=0.18, compared to mAP=0.25 on ZOD

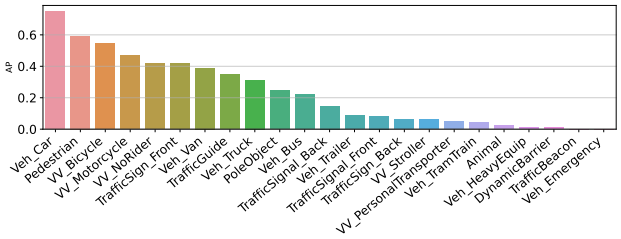


Figure 7: 3D object detection AP per class (0-250m).

consistently detected, nearly half of the analyzed classes register an AP score of under 10%.

In the realm of traffic sign recognition (TSR), a significant challenge emerges from the marked class imbalance between commonly recognized, universal signs (*e.g.*, pedestrian crossing) and those specific to rare scenarios (*e.g.*, warning for polar bears). We implement a simple ResNet-50-based classification baseline for TSR on ZOD and report the results in Table 5. We exclude classes with less than 10 signs in the validation set. Common signs are classified very well, as shown by high F1-micro score, whereas the class-balanced F1-macro score is significantly lower. We also show accuracy for the 10 most common and 10 least common classes, which further highlight the difference in performance on common vs. rare classes. Jointly, these results warrants further investigation into long-tailed tasks to enhance the reliability of autonomous systems.

4.4. Anonymization

To analyze the effects that anonymization has on downstream computer vision tasks, we train two image-based object detectors on three different versions of the images: the original, blurred, and DNAT images. Training is done on each of the anonymized image sets separately, while the evaluation is performed on the original images. We train on the three classes for which anonymization is relevant, namely vehicles, vulnerable vehicles (bicycles, motorcycles, wheelchairs, *etc*.), and pedestrians. The first detection pipeline is a Faster-RCNN [22], coupled with a feature pyramid network [15] and a Resnet-50 [13] backbone – implemented in the Detectron2 [31] framework – while the second pipeline is YOLOv7 [28]. The results are computed according to the COCO evaluation protocol [16] and presented in Table 3 (more comprehensive results are available in the supplementary material). The results show no statistically significant performance degradation when training with anonymized images over the original setting. Moreover, these results act as a baseline for image-based object detection on ZOD *Frames*.

5. Conclusion

We present ZOD, a diverse multimodal dataset for autonomous driving. ZOD contains data collected, with high-resolution sensors, from 14 European countries, thereby addressing the lack of European data in publicly available AD datasets. With this geographical diversity we are able to provide a wide range of driving scenarios, covering snowy country roads in northern Sweden, rainy highways in Germany, busy downtown traffic in France, and sunny suburban roads in Italy. We supply a comprehensive set of dense annotations, including semantic/instance segmentation masks for lane markings, road paintings, and ego road, 2D/3D bounding boxes for objects, and classification labels of the road surface condition. Notably, the 3D object annotations range up to 245 meters, which is farther than any comparable AD dataset. Additionally, ZOD boasts a comprehensive traffic sign taxonomy, with a greater number of unique instances than any other comparable datasets. In terms of future work, we will add temporal-consistent annotations for the entire duration of the sequences in ZOD *Sequences*, two seconds of supporting camera frames for all ZOD *Frames*, and more *Sequences* and *Drives* with supporting high definition maps. We hope our diverse dataset can inspire research that drives the field even further toward robust and safe AD.

6. Limitations

While we are excited about ZOD's potential, it's crucial to acknowledge its constraints. The dataset currently offers keyframe annotations, suitable for various tasks (including tasks that require spatiotemporal reasoning) but lacks support for consistent object tracking over time. We plan to enhance this with full sequence annotations. While we've analyzed the effect of anonymization on 2D object detection in ZOD, the impact on other tasks remains unexplored. Our 3D annotations reach up to 250 meters, but their quality and recall wane at these distances. This gap is bridged somewhat by our 2D annotations, which extend much further. Finally, annotations for traffic signs and ego-road are unavailable for some *Frames*. We look forward to improving ZOD with future updates and encourage community feedback.

7. Acknowledgements

The Zenseact Open Dataset would not be possible without the support of several individuals within our organization. We would like to thank, without any particular order, the following people for their contributions to the dataset: Oleksandr Panasenko, Jakub Bochynski, Dónal Scanlan, Benny Nilsson, Jonas Ekmark, Bolin Shao, Georgios Efthymiou, Erik Rosén, Oleksii Khakhlyuk, Pavel Lutskov, Maryam Fatemi, Joakim Johnander, Mats Nordlund, Joakim Frid, Goran Widborn, and Erik Coelingh.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] CC BY-SA 4.0 license page. https://creativecommons.org/licenses/by-sa/4.0/. Accessed: 2022-10-05.

[2] BrighterAI. BrighterAI web page. https://brighter.ai/. Accessed: 2022-10-05.

[3] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[5] Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[6] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The mapillary traffic sign dataset for detection and classification on a global scale. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 68–84. Springer, 2020.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[12] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Fredrik Larsson and Michael Felsberg. Using fourier descriptors and spatial models for traffic sign recognition. In *Image Analysis*, pages 238–249, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[15] 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*, pages 2117–2125, 2017.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer International Publishing, 2014.

[17] Jiageng Mao, Niu Minzhe, ChenHan Jiang, hanxue liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing XU, and Hang Xu. One million scenes for autonomous driving: Once dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[18] Markus Mathias, Radu Timofte, Rodrigo Benenson, and Luc Van Gool. Traffic sign recognition — how far are we from the solution? In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[19] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[20] World Health Organization. Global status report on road safety 2018. Technical report, World Health Organization, 2018.

[21] 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)*, pages 9552–9557, 2019.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[23] V.I. Shakhuro and Anton Konushin. Russian traffic sign images dataset. *Computer Optics*, 40:294–300, 01 2016.

[24] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

[25] 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*, pages 2446–2454, 2020.

[26] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3614–3633, 2022.

[27] Paul Voigt and Axel Von dem Bussche. *The eu general data protection regulation (GDPR)*. Springer, Cham, 2017.

[28] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[29] Peng Wang, Xinyu Huang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[30] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James 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.

[31] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[32] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021.

[33] 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*, pages 11784–11793, 2021.

[34] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[35] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 09 2017.

[36] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

[37] Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu. Traffic-sign detection and classification in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2110–2118, 2016.