Ad-datasets: A Meta-collection of Data Sets for Autonomous Driving

Daniel Bogdoll[1,2,*], Felix Schreyer[2,*] and J. Marius Zöllner[1,2]
[1]*FZI Research Center for Information Technology, Germany*
[2]*Karlsruhe Institute of Technology, Germany*

Keywords: Autonomous Driving, Data Set, Overview, Collection.

Abstract: Autonomous driving is among the largest domains in which deep learning has been fundamental for progress within the last years. The rise of datasets went hand in hand with this development. All the more striking is the fact that researchers do not have a tool available that provides a quick, comprehensive and up-to-date overview of data sets and their features in the domain of autonomous driving. In this paper, we present *ad-datasets*, an online tool that provides such an overview for more than 150 data sets. The tool enables users to sort and filter the data sets according to currently 16 different categories. *ad-datasets* is an open-source project with community contributions. It is in constant development, ensuring that the content stays up-to-date.

1 INTRODUCTION

One of the core building blocks on the way to fully autonomous vehicles are data sets. Of particular interest are those that contain data on all aspects of traffic. Their area of application in the research area related to autonomous driving is diverse. Hence, their number has multiplied significantly over the years. They have proven to be a necessary tool on the way to achieving the goal of autonomous vehicles.

Given this increase in importance, it seems all the more surprising that researchers still do not have a tool at hand, that provides them an overview of existing data sets and their characteristics. Even today, the search for fitting data sets is a tedious and cumbersome task. Existing overviews are typically either incomplete and miss relevant data sets or come in the form of scientific papers, therefore slowly but steadily becoming outdated. This is an especially undesirable condition in such a rapidly evolving field.

Researchers are therefore regularly reliant on finding suitable data sets on their own. However, this task is not only extremely time-consuming, as it involves studying numerous websites and papers. It also in no way guarantees that researchers will indeed find a suitable, perhaps even optimal, data set. This led to the situation that many researchers rely on the same, old datasets, while newer, larger options might be available.

In this publication, we present our attempt to address this unanswered challenge. With *ad-datasets*[1] we have developed a tool that provides users with a comprehensive, up-to-date overview of existing data sets in the field of autonomous driving, as shown in Figure 1. In addition, the properties of the data sets are broken down into different categories. Users are given the opportunity to individually filter and sort the data sets according to certain criteria or rather, to their individual needs. Currently[2], *ad-datasets* comprises 158 data sets.

Furthermore, *ad-datasets* is an open-source project. The community is encouraged to contribute, such as missing data sets or metadata. We are proud to have already received several community contributions only a few weeks after the initial release, including ones from ARGO AI, a well-known company within the domain. *ad-datasets*, hosted via Github Pages, enables this particularly easily via pull requests and automated deployments.

The further structure of this paper is organized as follows: In section 2, relevant related work is presented. Section 3 introduces the tool *ad-datasets* and its technical implementation. Section 4 analyzes the information obtained. The final section 5 provides an outlook into future extensions.

[*] These authors contributed equally

[1] https://ad-datasets.com/

[2] as of Nov 25, 2021

Bogdoll, D., Schreyer, F. and Zöllner, J.
Ad-datasets: A Meta-collection of Data Sets for Autonomous Driving.
DOI: 10.5220/0011001900003191
In *Proceedings of the 8th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2022)*, pages 46-56
ISBN: 978-989-758-573-9; ISSN: 2184-495X

2 RELATED WORK

The research area around autonomous driving shines with great progress and a high rate of development. Numerous advances are made every year and the amount of literature continues to grow. This is also the case in the area of data collection, respectively the area of data sets in autonomous driving. Over the years, however, the sheer amount of data sets has become increasingly complicated. For this reason, there have been attempts in the past to provide a structure to these advances. In general, these attempts can be divided into two separate categories.

2.1 Scientific Papers

First, there are studies that aim to create a general, comprehensive overview of existing data sets from the area of research. These include (Yin and Berger, 2017), where 27 data sets prevalent at the time were presented. Around three times that amount was summarized in (Laflamme et al., 2019).

In addition to these studies, which aim directly at creating an overview of data sets, there are also studies that focus primarily on different research questions, but also contain such an overview. (Feng et al., 2020) presented their work, which also comes in an online version briefly mentioned in the following subsection, that investigated the research question of *Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving*. Yet, the authors also examined multi-modal data sets, with the multimodality referring to the sensors used. Compared to the works of Yin and Berger and Laflamme et al. however, this collection is much smaller in scope. (Heidecker et al., 2021) takes on the topic of corner cases in highly automated driving. Here, too, a section revolves solely around suitable data sets for corner case detectors. Finally, in the same year (Kim and Hwang, 2021) was published that contains a survey addressing data sets for monocular 3D detection. Yet, these papers also lag behind in scope. It is important to note, that publications which do not focus on data sets but provide overviews of them as side effect typically focus strongly on their area of research. Thus, they rarely provide a broad picture of data sets and tend to focus either on popular and well known data sets or their specific niche.

What all these publications share is that they appeared in the format of scientific works. Thus, they have some weaknesses in common when it comes to searching for data sets. To start with, they share the problem that they become out-of-date relatively quickly. This is an undesirable characteristic, especially in a research area that is developing as quickly as the one of autonomous driving. In addition, these overviews lack a convenient format. Naturally, the publications offer neither a filtering nor a sorting function, therefore not being as effortless and time saving as one would desire.

2.2 Online Sources

Further, there exist various sources in online formats. These can in turn be broken down into read-only textual sources, mostly in the form of blogs (Choudhury, 2020)(Cambridge Spark, 2018)(Nguyen's, 2021), wiki entries (Wikipedia, 2021)(FOT-Net, 2020), git repositories (Heyuan, 2019)(Diaz, 2021) or miscellaneous entries (DIY Robocars, 2017)(Krunal, 2018)(Feng et al., 2021), and into interactive tools.

The textual sources are usually kept very compact and often contain fewer than ten data sets, which are typically among the better known ones. Additionally, many times the summaries are primarily aimed at machine learning related data sets in general. It is then left to the users to filter out the relevant autonomous driving related data sets. Since all of these sources, much like scientific papers, do neither allow any filtering nor sorting functions and are typically not kept up-to-date, they are poorly suited as sources for extensive data set searches.

A completely different picture emerges when looking at online tools. Their format is much more suitable for searches of any kind.

RList (RList, 2021) provides such a tool. Both sorting and search functions are available to users. Further, entries are broken down into categories by release date, organization, frames and location. However, with only nine data set entries it remains rather small. *Scale* (Scale, 2019) provides another, much more extensive tool. It comprises 50 data sets from the areas of autonomous driving and natural language processing. Further, the tool highlights a variety of categories, namely sensor types, annotations, diversity and recording location. Filter functions are also available to the user. Unfortunately, however, the tool does not contain any data set entries after 2019 and can therefore not be considered as up-to-date. This is where the *Dataset list* (Plesa, 2021) tool stands out. Up to the time of publication of this work, the tool was regularly updated, and can therefore be regarded as up-to-date. The tool comes with four categories of which the highlighting of the licenses differs from the ones previously described. Yet, with a volume of 25 data sets from the field of autonomous driving, it is

Table 1: Comparison between *ad-datasets* and other online tools. We compare the number of entries in each tool, the features the tools provide, the number of properties broken down in detail and when the tools were last updated.

Dataset Overview	Entries	Filterable	Sortable	Community Contribution	Number of Itemized Properties	Last Update
ad-datasets	**158**	**Yes**	**Yes**	**Yes**	**16**	**2021**
Bifrost	50	**Yes**	**Yes**	No	-	2020
Dataset list	25	**Yes**	No	**Yes**	4	**2021**
Kaggle	31	**Yes**	**Yes**	**Yes**	-	**2021**
RList	9	**Yes**	**Yes**	No	5	**2021**
Scale	50	**Yes**	No	No	4	2019
YonoStore	10	No	**Yes**	No	-	**2021**

too small to be able to claim a complete overview of existing data sets.

Also *Bifrost* with 50 data sets (Bifrost, 2020), *Kaggle* with 31 data sets (Kaggle, 2021) and *YonoStore* with ten data sets (YonoStore, 2021) cannot be seen as complete, either. Further, unlike the aforementioned online tools, they do not provide the user with an overview of data set properties at first glance.

More general offers such as *Google Dataset Search* (Google, 2018), *DeepAI Datasets* (DeepAI, 2017) or *Papers with Code Datasets* (Facebook AI, 2018) have the issue that no domain-specific overview is possible. Therefore, although they list numerous data sets, they are not suitable.

Finally, it should be noted that the majority of the tools examined here do not allow the user to contribute content. Only *Kaggle* and *Dataset list* provide the community with an opportunity to independently add missing data sets.

3 SELECTION PROCESS

ad-datasets is an online tool designed as the central point of contact for data sets in the field of autonomous driving. It includes a detailed representation of the data sets according to 16 different property categories and enables users to interact via filter and sorting functions. As of writing, *ad-datasets* comprises 158 data sets, 40 of which were examined in detail according to the 16 categories.

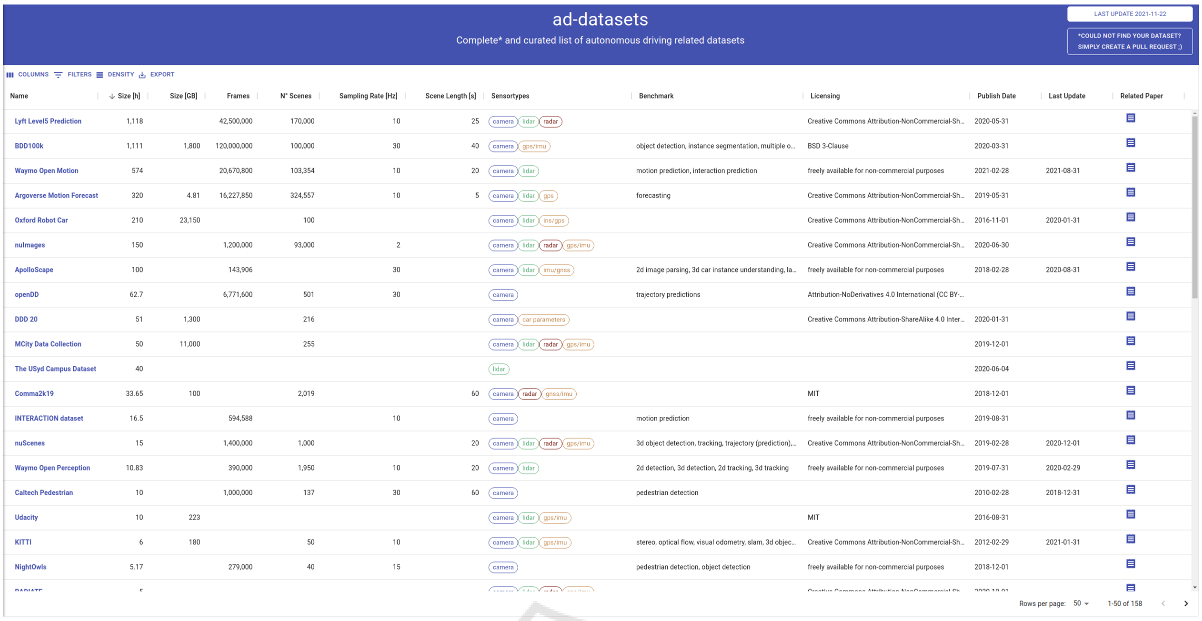
3.1 Content

The search for data sets poses a major challenge. In this work, large and well-known data sets were easily found, both via the numerous online sources and the numerous papers which refer to them. Less known, older data sets were collected through an extensive literature research. An approach that has proven to be well-suited for their finding has been the snowballing principle. However, newer, lesser-known publications cannot be found this way. In fact, finding them has turned out to be the greatest difficulty. Since they are rarely mentioned in literature or online sources, their search had to be designed differently. In this work, these data sets were identified through the use of search engines and via communities such as LinkedIn. It should be noted, that this procedure is associated with a high level of effort, so being well-connected in the relevant communities is of great benefit. Needless to say, this practice is not very scientific, but extremely effective, as new data sets in the community are often shared through social media.

In this paper, autonomous driving related data sets are defined as data sets that contain data on all aspects of traffic. They can both consist of scenes and scenarios[3] of road traffic or its participants. *ad-datasets* includes, for example, data sets with video sequences of intersections from a bird's eye view, but also recordings from a vehicle directly participating in traffic.

The primary focus of *ad-datasets* are those data sets which were published after the famous KITTI data set, which serves, so to speak, as a time benchmark. Regarding the selection of the data sets which were analyzed in detail, the selection can be separated into two parts. 31 of the data sets were selected manually by the authors, focusing on the most popular ones. The remaining nine data sets were selected at random. The breakdown can be seen in table 2.

[3] Following the definitions of (Ulbrich et al., 2015)



ad-datasets

Complete* and curated list of autonomous driving related datasets

LAST UPDATE 2021-11-22
*COULD NOT FIND YOUR DATASET? SIMPLY CREATE A PULL REQUEST ;)
COLUMNS
FILTERS
DENSITY
EXPORT

Name	↓ Size [h]	Size [GB]	Frames	N° Scenes	Sampling Rate [Hz]	Scene Length [s]	Sensortypes	Benchmark	Licensing	Publish Date	Last Update	Related Paper
Lyft Level5 Prediction	1,118		42,500,000	170,000	10	25	camera lidar radar		Creative Commons Attribution-NonCommercial-Sh...	2020-05-31		▤
BDD100k	1,111	1,800	120,000,000	100,000	30	40	camera gps/imu	object detection, instance segmentation, multiple o...	BSD 3-Clause	2020-03-31		▤
Waymo Open Motion	574		20,670,800	103,354	10	20	camera lidar	motion prediction, interaction prediction	freely available for non-commercial purposes	2021-02-28	2021-08-31	▤
Argoverse Motion Forecast	320	4.81	16,227,850	324,557	10	5	camera lidar gps	forecasting	Creative Commons Attribution-NonCommercial-Sh...	2019-05-31		▤
Oxford Robot Car	210	23,150		100			camera lidar ins/gps		Creative Commons Attribution-NonCommercial-Sh...	2016-11-01	2020-01-31	▤
nuImages	150		1,200,000	93,000	2		camera lidar radar gps/imu		Creative Commons Attribution-NonCommercial-Sh...	2020-06-30		▤
ApolloScape	100		143,906		30		camera lidar imu/gnss	2d image parsing, 3d car instance understanding, la...	freely available for non-commercial purposes	2018-02-28	2020-08-31	▤
openDD	62.7		6,771,600	501	30		camera	trajectory predictions	Attribution-NoDerivatives 4.0 International (CC BY-...			▤
DDD 20	51	1,300		216			camera car parameters		Creative Commons Attribution-ShareAlike 4.0 Inter...	2020-01-31		▤
MCity Data Collection	50	11,000		255			camera lidar radar gps/imu			2019-12-01		▤
The USyd Campus Dataset	40						lidar			2020-06-04		▤
Comma2k19	33.65	100		2,019		60	camera radar gnss/imu		MIT	2018-12-01		▤
INTERACTION dataset	16.5		594,588		10		camera	motion prediction	freely available for non-commercial purposes	2019-08-31		▤
nuScenes	15		1,400,000	1,000		20	camera lidar radar gps/imu	3d object detection, tracking, trajectory (prediction),...	Creative Commons Attribution-NonCommercial-Sh...	2019-02-28	2020-12-01	▤
Waymo Open Perception	10.83		390,000	1,950	10	20	camera lidar	2d detection, 3d detection, 2d tracking, 3d tracking	freely available for non-commercial purposes	2019-07-31	2020-02-29	▤
Caltech Pedestrian	10		1,000,000	137	30	60	camera	pedestrian detection		2010-02-28	2018-12-31	▤
Udacity	10	223					camera lidar gps/imu		MIT	2016-08-31		▤
KITTI	6	180		50	10		camera lidar gps/imu	stereo, optical flow, visual odometry, slam, 3d objec...	Creative Commons Attribution-NonCommercial-Sh...	2012-02-29	2021-01-31	▤
NightOwls	5.17		279,000	40	15		camera	pedestrian detection, object detection	freely available for non-commercial purposes	2018-12-01		▤
RADIATE	5						camera lidar radar gps/imu		Creative Commons Attribution-NonCommercial-Sh...	2020-10-01		▤

Rows per page: 50 ▾ 1-50 of 158 ‹ ›

Figure 1: Screenshot of the ad-datasets web application.

3.2 Structure

The selected property categories are in turn the result of an expert survey conducted of the research group for technical cognitive systems at the FZI Research Center for Information Technology. Over 20 different categories were suggested in the survey (Table 3). Ultimately, 16 categories found their way into the initial version due to time constraints. The selection of these properties that were included in the tool was made based on an examination of ten exemplary data sets. In this examination, the time required to collect the data for each property was investigated. The data was collected via the web presences of the data sets and their corresponding papers. The decision on whether a category was included was made based upon the author's perceived importance of the property and the associated time required to include the category in the selection. The ten exemplary data sets were Cityscapes 3D, ApolloScape, Lyft Level5 Prediction, Oxford Robot Car, nuScenes, PandaSet, Waymo Open Motion, KITTI, BDD100k and openDD. The resulting 16 categories are presented in detail at this point.

3.2.1 Annotations

Annotations. This property describes the types of annotations with which the data sets have been provided.

Benchmark. If benchmark challenges are explicitly listed with the data sets, they are specified here.

Frames. Frames states the number of frames in the data set. This includes training, test and validation data.

Last Update. If information has been provided on updates and their dates, they can be found in this category.

Licensing. In order to give the users an impression of the licenses of the data sets, information on them is already included in the tool.

Location. This category lists the areas where the data sets have been recorded.

N° Scenes. N° Scenes shows the number of scenes contained in the data set and includes the training, testing and validation segments. In the case of video recordings, one recording corresponds to one scene. For data sets consisting of photos, a photo is the equivalent to a scene.

Publish Date. The initial publication date of the data set can be found under this category. If no explicit information on the date of publication of the data set could be found, the submission date of the paper related to the set was used at this point.

Related Data Sets. If data sets are related, the names of the related sets can be examined as well. Related data sets are, for example, those published by the same authors and building on one another.

Related Paper. This property solely consists of a link to the paper related to the data set.

Sampling Rate [Hz]. The Sampling Rate [Hz] property specifies the sampling rate in Hertz at which the sensors in the data set work. However, this declaration is only made if all sensors are working at

Table 2: Overview of the data sets analyzed in detail. Data sets marked with ^{j} belong to those which were chosen randomly, data sets marked with ^{i} have been selected deterministically by the authors.

Name	Size [h]	Frames	N° Scenes	Sensortypes				Publish Date	Sources
				camera	lidar	radar	other		
ApolloScape^{i}	100	143,906		✓	✓		✓	2018.03	(Wang et al., 2019)(ApolloScape, 2018)
Argoverse Motion Forecasting^{i}	320		324,557	✓	✓		✓	2019.06	(Chang et al., 2019)(ARGO AI, 2019)
Argoverse 3D Tracking^{i}			113	✓	✓		✓	2019.06	(Chang et al., 2019)(ARGO AI, 2019)
A2D2^{i}		433,833	3	✓	✓		✓	2020.04	(Geyer et al., 2020)(A2D2, 2020)
BDD100k^{i}	1,111	120,000,000	100,000	✓			✓	2020.04	(Yu et al., 2020)(ETH VIS Group, 2018)
Bosch Small Traffic Lights^{j}		13,427		✓				2017.05	(Behrendt and Novak, 2017)(for Image Processing, 2017)
Caltech Pedestrian^{i}	10	1,000,000	137	✓				2010.03	(Dollár et al., 2009)(California Institute of Technology, 2009)
Cityscapes 3D^{i}				✓			✓	2016.02	(Cordts et al., 2016)(Gählert et al., 2020)(Cityscapes Dataset, 2016)
Comma2k19^{i}	33.65		2,019	✓		✓	✓	2018.12	(Schafer et al., 2018)(comma.ai, 2019)
DDD 20^{j}	51		216	✓			✓	2020.02	(Hu et al., 2020)(Inst. of Neuroinformatics, Univ. of Zurich and ETH Zurich, 2020)
Fishyscapes^{i}				✓				2019.09	(Blum et al., 2019)(ETH Zürich, 2019)
Ford Autonomous Vehicle^{i}				✓	✓		✓	2020.03	(Agarwal et al., 2020)(Ford, 2020)
H3D^{j}	0.77	27,721	160	✓	✓		✓	2019.03	(Patil et al., 2019)
India Driving^{j}		10,004	182	✓				2018.11	(Varma et al., 2019)(INSAAN, 2018)
INTERACTION^{i}	16.5	594,588		✓				2019.09	(Zhan et al., 2019)(INTERACTION Dataset Consortium, 2019)
KAIST Multi-Spectral Day/Night^{i}				✓	✓		✓	2017.12	(Choi et al., 2018)(KAIST, 2017)
KAIST Urban^{i}			18	✓	✓		✓	2017.09	(Jeong et al., 2019)(Autonomy and Lab, 2021)
KITTI^{i}	6		50	✓	✓		✓	2012.03	(Geiger et al., 2013)(cvlibs, 2012)
KITTI-360^{i}		400,000		✓	✓		✓	2015.11	(Xie et al., 2016)(cvlibs, 2021)
LostAndFound^{i}		21,040	112	✓				2016.09	(Pinggera et al., 2016)(6D-Vision, 2016)
Lyft Level5 Perception^{i}	2.5		366	✓	✓			2019.07	(Houston et al., 2020)(Level 5, 2020a)
Lyft Level5 Prediction^{i}	1,118	42,500,000	170,000	✓	✓	✓		2020.06	(Houston et al., 2020)(Level 5, 2020b)
MCity Data Collection^{i}	50		255	✓	✓	✓	✓	2019.12	(Dong et al., 2019)
NightOwls^{j}	5.17	279,000	40	✓				2018.12	(Neumann et al., 2018)(NightOwls dataset, 2018)
nuImages^{i}	150	1,200,000	93,000	✓	✓	✓	✓	2020.07	(Motional, 2019a)
nuScenes^{i}	15	1,400,000	1,000	✓	✓	✓	✓	2019.03	(Caesar et al., 2019)(Motional, 2019b)
openDD^{i}	62.7	6,771,600	501	✓				2020.09	(Breuer et al., 2020)(L3Pilot, 2019)
Oxford Radar Robot Car^{i}			32	✓	✓	✓	✓	2020.02	(Barnes et al., 2020)(Oxford Robotics Institute, 2020)
Oxford Robot Car^{i}	210		100	✓	✓		✓	2016.11	(Maddern et al., 2017)(Oxford Robotics Institute, 2017)
PandaSet^{i}	0.23	48,000	103	✓	✓		✓	2020.04	(Hesai, Scale AI, 2020)
RadarScenes^{j}	4		158	✓		✓	✓	2021.03	(Schumann et al., 2021)(RadarScenes, 2021)
RADIATE^{j}	5			✓	✓	✓	✓	2020.10	(Sheeny et al., 2020)(Vision Lab, Perception and Robotics Group, 2020)
RoadAnomaly21^{i}		100	100	✓				2021.04	(Chan et al., 2021)(Segment Me If You Can, 2021)
Semantic KITTI^{i}		43,552	21		✓			2019.07	(Behley et al., 2019)(University of Bonn, 2019)
Synscapes^{j}		25,000	25,000	✓				2018.10	(Wrenninge and Unger, 2018)(7D Labs Inc., 2018)
Udacity^{i}	10			✓	✓		✓	2016.09	(Udacity, 2020)
Waymo Open Motion^{i}	574	20,670,800	103,354	✓	✓			2021.03	(Ettinger et al., 2021)(Waymo LLC, 2021)
Waymo Open Perception^{i}	10.83	390,000	1,950	✓	✓			2019.08	(Sun et al., 2020)(Waymo LLC, 2019)
WildDash^{i}			156	✓				2018.02	(Zendel et al., 2018)(AIT, 2020)
4Seasons^{j}			30	✓			✓	2020.10	(Wenzel et al., 2020)(Artisense, 2020)

the same rate or, alternatively, if the sensors are being synchronized. Otherwise, the field remains empty.

Scene Length [s]. This property describes the length of the scenes in seconds in the data set, provided all scenes have the same length. Otherwise, no information is given. For example, if a data set has scenes with lengths between 30 and 60 seconds, no entry can be made. The background to this procedure is to maintain comparability and sortability.

Sensor Types. This category contains a rough description of the sensor types used. Sensor types are, for example, lidar or radar.

Sensors - Details. The Sensors - Detail category is an extension of the Sensor Types category. It includes a more detailed description of the sensors. The sensors are described in detail in terms of type and number, the frame rates they work with, the resolutions which sensors have and the horizontal field of view.

Size [GB]. The category Size [GB] describes the storage size of the data set in gigabytes.

Size [h]. The Size [h] property is the equivalent of the Size [GB] described above, but provides information on the size of the data set in hours.

It should be noted that, the name of the data set is naturally listed, too. It further acts as a link to the respective website of the data set. Further, it is worth mentioning that the general aim is to state the properties as precisely as possible. Yet, this also depends on how accurate the documentation of the data set is. For example, a size specification of 100+ hours is less precise than a specification of 103 hours. In the case of the former, the tool would only state a size of 100 hours.

3.3 Technical Implementation

ad-datasets is hosted on GitHub Pages. This allows for a seamless integration of the research community. Not only can the community influence the development process, but also actively expand its content. This is especially valuable as, no matter how extensive a search, it cannot be guaranteed that the tool is indeed complete. Missing entries or metadata can be added via pull requests, which automatically trigger the build system to publish the changes.

The implementation of the tool itself was done using the frameworks React (Facebook, 2013) and Material-UI (Material-UI, 2014). The latter enables a quick and uncomplicated creation of filterable as well as sortable tables. In this work, the data grid component (Material-UI, 2020) of the Material-UI framework was used.

Table 3: Property categories resulting from an expert survey with indication whether they have been included.

Property	Included
Annotations	✓
Benchmark	✓
Frames	✓
Last Update	✓
Licensing	✓
Location	✓
N° Scenes	✓
Publish Date	✓
Related Datasets	✓
Related Paper	✓
Sampling Rate [Hz]	✓
Scene Length [s]	✓
Sensor Types	✓
Sensors - Details	✓
Size [GB]	✓
Size [h]	✓
Data Format	✗
Main Focus	✗
Recording Perspective	✗
Statistics	✗
Tooling	✗

4 PROPERTY EXAMINATION

This section takes a closer look at the 40 data sets from *ad-datasets* that have been analyzed in detail.

4.1 Data Sets over Time

In Figure 2, the observed data sets are shown on a timeline. While keeping in mind that the majority of the data sets were not chosen randomly, a clear tendency can be identified that both the amount of data sets and the amount of data set publications per year are steadily increasing.

For a start, this finding emphasizes once again the dynamics in the research field of autonomous driving. More to the point, it demonstrates the importance of a well-maintained and up-to-date tool in order to tackle the currently prevailing inconvenience that come with data set searches. The accelerating pace of data set publications makes it even clearer, that unsupervised overviews age at an equally increasing pace.

4.2 Use of Sensor Types

When looking at the sensor types used in the various data sets (Figure 3), it is first noticeable that the sensor type used most frequently is the camera. In fact, the only data set analyzed that does not utilize camera sensors is the Semantic KITTI data set. The remaining 39 out of 40 data sets make use of the camera sensor type.

The Semantic KITTI data set in return deploys lidar sensors. 22 other data sets do the same, so that more than 50% of the data sets include lidar data. Radar data is used much less frequently. Only eight data sets make use of the sensors. When additionally considering publication dates, it shows that radar data sets were added much later. None of the eight radar data sets have been published before the end of 2018. Yet, it seems that the importance of data sets containing radar data is increasing. Of the ten data sets published in 2020, four contained radar data.

The outsiders among the sensor types include thermal cameras and thermometers. Each appear in only one data set, the thermal camera in Cityscapes 3D and the thermometer in the Multi-Spectral Day / Night data set.

4.3 Size

Upon analyzing the size of data sets, one can distinguish between storage size, the time span and the number of scenes of the data sets.

The inspection of storage size (Figure 5) reveals that of the eleven data sets which provide information on their size, five data sets are smaller than 1,000 GB with the median being 1,300 GB. At the same time, however, there coexist significantly larger data sets. The MCity Data Collection data set is 11,000 GB large. The even larger Oxford Robot Car set is in fact with 23,150 GB more than 17 times as large as the median. It can therefore be seen that data sets are mainly of a similar order of magnitude in terms of storage size. However, there are also sets of much larger sizes. When looking at the scope of time (Table 4), an even more pronounced picture emerges. The median of the 23 data sets, for which the corresponding information could be obtained, is 16.5 hours. There are more data sets smaller than ten hours than there are data sets larger than 100 hours. But here, too, there are examples of very large data sets. The Lyft Level5 Prediction data set is 1,118 hours large, the BDD100k data set 1,111 hours. These two data sets are over 60 times as large as the median.

On closer inspection, however, one can spot a difference between the storage wise large data sets and

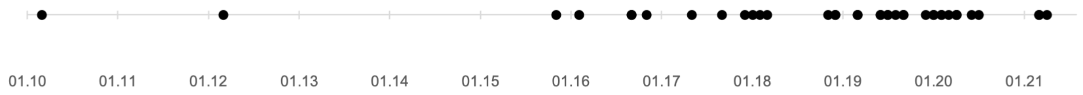


Figure 2: Timeline when data sets have been published.

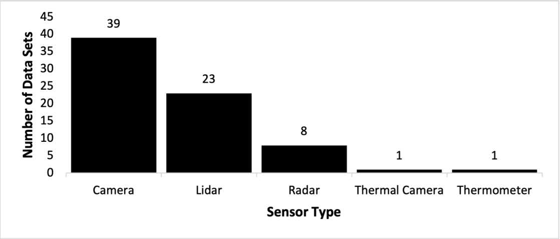


Figure 3: Frequency of use of sensor types in the data sets.

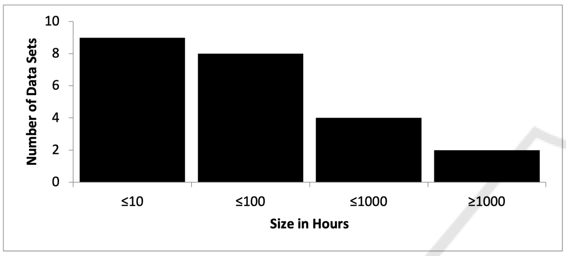


Figure 4: Histogram depicting the distribution of the data sets over their scope of time.

the time wise large data set. Some of those data sets that are rather large in terms of storage space have been existing for multiple years. For example, the Oxford Robot Car data set was published back in 2016. Data sets which feature large scopes of time have been published rather recently. Lyft Level5 Prediction and BDD100k had not been published until 2020.

These impressions can also be transferred to the analysis of the number of scenes in data sets (Table 6). Once again, the majority revolves around a similar range regarding the number of scenes. However, again, there are outliers that are significantly larger. The median over all 30 data sets, for which information could be gathered, is 158 scenes. Opposite to that, the Lyft Level5 Prediction (170,000 scenes) and the Argoverse Motion Forecasting (324,557 scenes) are over 1,000 respectively over 2,000 times as large.

What's more, data sets that come with large number of scenes have been published relatively recently as well.

5 CONCLUSION AND OUTLOOK

With *ad-datasets*, we have presented a tool that aims to simplify the previously complex and time-consuming search for suitable data sets related to the research area of autonomous driving. *ad-datasets* offers users an overview of over 150 data sets, which are further broken down into 16 different properties. Finally, users can interact with the tool using filter and sorting functions. The timeliness of *ad-datasets* is maintained through further maintenance of the tool and can be supported by the community, e.g., via pull requests.

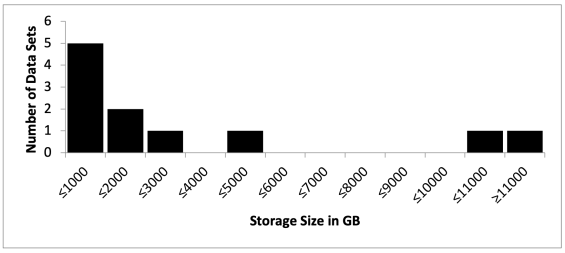


Figure 5: Histogram depicting the distribution of the data sets over their storage size.

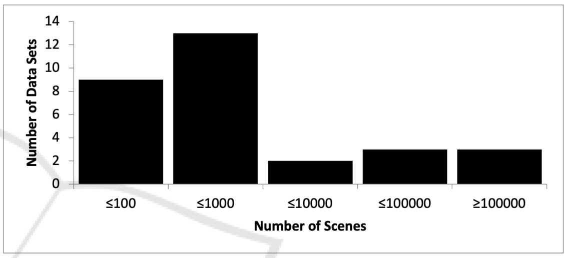


Figure 6: Histogram depicting the distribution of the data sets over their number of scenes.

However, the full potential of the tool has not yet been realized. To fully do so, a couple of aspects can be addressed. First, it is obviously necessary to complete the detailed analysis of the remaining data sets. This is already an ongoing work in progress.

In addition, it is crucial to obtain feedback from the research community. After all, they are the target audience of *ad-datasets*. Hence, their feedback is essential in developing a truly value-adding tool. Initial feedback from the community has been very positive and has already led to contributions.

Finally, it must be borne in mind that some of the properties suggested in the expert survey did not find their way into the initial version of *ad-datasets*. Therefore, they are subject to future work. It was proposed to include information on the statistical distribution of classes, labels etc. For the time being, this proposal remains subject to future work, as there were concerns regarding copyright. Likewise, the suggested property categories data format and tooling options remain subject to future work. Important categories, which for the moment were associated with too much effort as well, are the recording perspective and the key aspects addressed by the data sets. In a later version of the tool, these will be included.

ACKNOWLEDGMENT

This work results partly from the KIGLIS project supported by the German Federal Ministry of Education and Research (BMBF), grant number 16KIS1231.

REFERENCES

6D-Vision (2016). LostAndFoundDataset. https://www.6d-vision.com/current-research/lostandfounddataset. Accessed 2022-02-03.

7D Labs Inc. (2018). Synscapes. https://7dlabs.com/synscapes-overview. Accessed 2021-08-12.

A2D2 (2020). Driving dataset. . Accessed 2021-08-22.

Agarwal, S., Vora, A., Pandey, G., Williams, W., Kourous, H., and McBride, J. (2020). Ford multi-AV seasonal dataset.

AIT (2020). Wilddash 2 benchmark. https://wilddash.cc/. Accessed 2021-08-09.

ApolloScape (2018). ApolloScape. http://apolloscape.auto/. Accessed 2021-08-04.

ARGO AI (2019). Argoverse. https://www.argoverse.org/. Accessed 2021-08-22.

Artisense (2020). 4seasons dataset. https://www.4seasons-dataset.com/. Accessed 2021-08-16.

Autonomy, I. R. and Lab, P. I. (2021). Complex urban dataset. https://sites.google.com/view/complex-urban-dataset.

Barnes, D., Gadd, M., Murcutt, P., Newman, P., and Posner, I. (2020). The oxford radar RobotCar dataset: A radar extension to the oxford RobotCar dataset. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., and Gall, J. (2019). SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*.

Behrendt, K. and Novak, L. (2017). A deep learning approach to traffic lights: Detection, tracking, and classification. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*.

Bifrost (2020). Search for visual datasets. https://datasets.bifrost.ai/. Accessed 2021-08-16.

Blum, H., Sarlin, P., Nieto, J. I., Siegwart, R., and Cadena, C. (2019). The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *arXiv preprint:1904.03215*.

Breuer, A., Termöhlen, J., Homoceanu, S., and Fingscheidt, T. (2020). opendd: A large-scale roundabout drone dataset. *arXiv preprint:2007.08463*.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2019). nuScenes: A multimodal dataset for autonomous driving. *arXiv preprint:1903.11027*.

California Institute of Technology (2009). Caltech Pedestrian Detection Benchmark. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/. Accessed 2021-08-22.

Cambridge Spark (2018). 50 free machine learning datasets: Self-driving cars. https://blog.cambridgespark.com/50-free-machine-learning-datasets-self-driving-cars-d37be5a96b28. Accessed 2021-08-16.

Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Salzmann, M., Fua, P., and Rottmann, M. (2021). SegmentMeIfYouCan: A benchmark for anomaly segmentation. *arXiv preprint:2104.14812*.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., and Hays, J. (2019). Argoverse: 3d tracking and forecasting with rich maps. *arXiv preprint:1911.02620*.

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J. S., An, K., and Kweon, I. S. (2018). Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3).

Choudhury, A. (2020). Top 10 popular datasets for autonomous driving projects. https://analyticsindiamag.com/top-10-popular-datasets-for-autonomous-driving-projects/. Accessed 2021-08-18.

Cityscapes Dataset (2016). Cityscapes dataset. https://www.cityscapes-dataset.com/. Accessed 2021-08-20.

comma.ai (2019). commai/comma2k19. https://github.com/commaai/comma2k19. Accessed 2021-08-10.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *arXiv preprint:1604.01685*.

cvlibs (2012). The kitti vision benchmark suite. http://www.cvlibs.net/datasets/kitti/. Accessed 2021-08-07.

cvlibs (2021). Kitti-360: A large-scale dataset with 3d&2d annotations. http://www.cvlibs.net/datasets/kitti-360/. Accessed 2021-08-21.

DeepAI (2017). Discover datasets for machine learning and a.i. https://deepai.org/datasets. Accessed 2021-08-20.

Diaz, M. (2021). manfreddiaz/awesome-autonomous-vehicles. https://github.com/manfreddiaz/awesome-autonomous-vehicles.

DIY Robocars (2017). Open datasets. https://diyrobocars.com/open-datasets/. Accessed 2021-08-20.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Dong, Y., Zhong, Y., Yu, W., Zhu, M., Lu, P., Fang, Y., Hong, J., and Peng, H. (2019). Mcity data collection for automated vehicles study. *arXiv preprint:1912.06258*.

ETH VIS Group (2018). Bdd100k. https://www.bdd100k.com/. Accessed 2021-08-07.

ETH Zürich (2019). The fishyscapes benchmark. https://fishyscapes.com/. Accessed 2021-08-21.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, A., Shlens, J., and Anguelov, D. (2021). Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Facebook (2013). React. https://reactjs.org/. Accessed 2021-08-22.

Facebook AI (2018). Datasets. https://paperswithcode.com/datasets. Accessed 2021-08-16.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3).

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. (2021). Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. https://boschresearch.github.io/multimodalperception/dataset.html. Accessed 2021-08-22.

for Image Processing, H. C. (2017). Bosch Small Traffic Lights Dataset. https://hci.iwr.uni-heidelberg.de/content/bosch-small-traffic-lights-dataset. Accessed 2021-08-12.

Ford (2020). Ford Autonomous Vehicle Dataset. https://avdata.ford.com/. Accessed 2021-08-20.

FOT-Net (2020). Automated driving datasets. https://wiki.fot-net.eu/index.php/Automated_Driving_Datasets. Accessed 2021-08-20.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A. S., Hauswald, L., Pham, V. H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., and Schuberth, P. (2020). A2d2: Audi autonomous driving dataset. *arXiv preprint:2004.06320*.

Gählert, N., Jourdan, N., Cordts, M., Franke, U., and Denzler, J. (2020). Cityscapes 3d: Dataset and benchmark for 9 DoF vehicle detection. *arXiv preprint:2006.07864*.

Google (2018). Datasetsearch - autonomous driving. https://datasetsearch.research.google.com/search?query=Autonomous%20driving&docid=L2cvMTFwd2Y0amZ0Yw%3D%3D. Accessed 2021-08-12.

Heidecker, F., Breitenstein, J., Rösch, K., Löhdefink, J., Bieshaar, M., Stiller, C., Fingscheidt, T., and Sick, B. (2021). An application-driven conceptualization of corner cases for perception in highly automated driving. *arXiv preprin:2103.03678*.

Hesai, Scale AI (2020). PandaSet by Hesai and Scale AI. https://pandaset.org/. Accessed 2021-08-05.

Heyuan, L. (2019). lhyfst/awesome-autonomous-driving-datasets. https://github.com/lhyfst/awesome-autonomous-driving-datasets. Accessed 2021-08-10.

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Jain, A., Omari, S., Iglovikov, V., and Ondruska, P. (2020). One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint:2006.14480*.

Hu, Y., Binas, J., Neil, D., Liu, S.-C., and Delbrück, T. (2020). DDD20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. *arXiv preprint:2005.08605*.

INSAAN (2018). India driving dataset. https://idd.insaan.iiit.ac.in/. Accessed 2021-08-12.

Inst. of Neuroinformatics, Univ. of Zurich and ETH Zurich (2020). DDD20: end-to-end DAVIS driving dataset. https://sites.google.com/view/davis-driving-dataset-2020/home.

INTERACTION Dataset Consortium (2019). INTERACTION Dataset. https://interaction-dataset.com/. Accessed 2021-08-18.

Jeong, J., Cho, Y., Shin, Y.-S., Roh, H., and Kim, A. (2019). Complex urban dataset with multi-level sensors from highly diverse urban environments. *The International Journal of Robotics Research*, 38(6).

Kaggle (2021). Datasets - autonomous driving. https://www.kaggle.com/datasets?search=Autonomous+Driving&sort=updated. Accessed 2021-08-15.

KAIST (2017). Visual Perception for Autonomous Driving. https://sites.google.com/view/multispectral. Accessed 2021-08-22.

Kim, S.-h. and Hwang, Y. (2021). A survey on deep learning based methods and datasets for monocular 3d object detection. *Electronics*, 10(4).

Krunal (2018). Semantic segmentation datasets for urban driving scenes. https://autonomous-driving.org/2018/07/15/semantic-segmentation-datasets-for-urban-driving-scenes/. Accessed 2021-08-22.

L3Pilot (2019). Opendd. https://l3pilot.eu/data/opendd. Accessed 2021-08-09.

Laflamme, C.-É. N., Pomerleau, F., and Giguere, P. (2019). Driving datasets literature review. *arXiv preprint:1910.11968*.

Level 5 (2020a). Perception dataset. https://level-5.global/data/perception/. Accessed 2021-08-20.

Level 5 (2020b). Prediction dataset. https://level-5.global/data/prediction/. Accessed 2021-08-20.

Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2017). 1 year, 1000km: The oxford RobotCar dataset. *The International Journal of Robotics Research (IJRR)*, 36(1).

Material-UI (2014). The React UI library you always wanted. https://material-ui.com/. Accessed 2021-08-22.

Material-UI (2020). Data grid. https://material-ui.com/components/data-grid/. Accessed 2021-08-22.

Motional (2019a). nuImages. https://www.nuscenes.org/nuimages. Accessed 2021-08-20.

Motional (2019b). nuScenes. https://www.nuscenes.org/. Accessed 2021-08-05.

Neumann, L., Karg, M., Zhang, S., Scharfenberger, C., Piegert, E., Mistr, S., Prokofyeva, O., Thiel, R., Vedaldi, A., Zisserman, A., and Schiele, B. (2018). NightOwls: A pedestrians at night dataset. In *Asian Conference on Computer Vision*.

Nguyen's, T. (2021). List of public large-scale datasets for autonomous driving research. https://tin.ng/public-datasets-for-autonomous-driving-research/. Accessed 2021-08-22.

NightOwls dataset (2018). NightOwls dataset. https://www.nightowls-dataset.org/. Accessed 2021-08-18.

Oxford Robotics Institute (2017). Oxford robotcar dataset. https://robotcar-dataset.robots.ox.ac.uk/. Accessed 2021-08-05.

Oxford Robotics Institute (2020). Oxford Radar RobotCar Dataset. https://oxford-robotics-institute.github.io/radar-robotcar-dataset/. Accessed 2021-08-18.

Patil, A., Malla, S., Gang, H., and Chen, Y.-T. (2019). The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. *arXiv preprint:1903.01568*.

Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., and Mester, R. (2016). Lost and found: Detecting small road hazards for self-driving vehicles. *arXiv preprint:1609.04653*.

Plesa, N. (2021). Machine learning datasets. https://www.datasetlist.com/. Accessed 2021-08-14.

RadarScenes (2021). RadarScenes. https://radar-scenes.com/. Accessed 2021-08-16.

RList (2021). List of autonomous driving open datasets. https://rlist.io/l/list-of-autonomous-driving-open-datasets?utm_source=insights.rlist.io&utm_medium=referral. Accessed 2021-08-16.

Scale (2019). Open datasets. https://scale.com/open-datasets. Accessed 2021-08-12.

Schafer, H., Santana, E., Haden, A., and Biasini, R. (2018). A commute in data: The comma2k19 dataset.

Schumann, O., Hahn, M., Scheiner, N., Weishaupt, F., Tilly, J. F., Dickmann, J., and Wöhler, C. (2021). RadarScenes: A real-world radar point cloud data set for automotive applications. *arXiv preprint:2104.02493*.

Segment Me If You Can (2021). Datasets. https://segmentmeifyoucan.com/datasets. Accessed 2021-08-09.

Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S., and Wallace, A. (2020). RADIATE: A radar dataset for automotive perception. *arXiv preprint:2010.09076*.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Udacity (2020). udacity/self-driving-car. https://github.com/udacity/self-driving-car/. Accessed 2021-08-22.

Ulbrich, S., Menzel, T., Reschka, A., Schuldt, F., and Maurer, M. (2015). Defining and substantiating the terms scene, situation, and scenario for automated driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*.

University of Bonn (2019). SemanticKITTI. http://www.semantic-kitti.org/. Accessed 2021-08-22.

Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., and Jawahar, C. (2019). Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Vision Lab, Perception and Robotics Group (2020). Heriot-Watt RADIATE Dataset. http://pro.hw.ac.uk/radiate/. Accessed 2021-08-12.

Wang, P., Huang, X., Cheng, X., Zhou, D., Geng, Q., and Yang, R. (2019). The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*.

Waymo LLC (2019). WAYMO Open Dataset - Perception. https://waymo.com/open/data/perception/. Accessed 2021-08-07.

Waymo LLC (2021). WAYMO Open Dataset - Motion. https://waymo.com/open/data/motion/. Accessed 2021-08-07.

Wenzel, P., Wang, R., Yang, N., Cheng, Q., Khan, Q., Stumberg, L. v., Zeller, N., and Cremers, D. (2020). 4seasons: A cross-season dataset for multi-weather SLAM in autonomous driving. *arXiv preprint:2009.06364*.

Wikipedia (2021). List of datasets for machine-learning research. https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research#Object_detection_and_recognition. Accessed 2021-08-15.

Wrenninge, M. and Unger, J. (2018). Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint:1810.08705*.

Xie, J., Kiefel, M., Sun, M.-T., and Geiger, A. (2016). Semantic instance annotation of street scenes by 3d to 2d label transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yin, H. and Berger, C. (2017). When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets. In *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*.

YonoStore (2021). Datasets. https://store.yonohub.com/product-category/datasets/?filters=product_cat[19]. Accessed 2021-08-12.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zendel, O., Honauer, K., Murschitz, M., Steininger, D., and Dominguez, G. F. (2018). WildDash - creating hazard-aware benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhan, W., Sun, L., Wang, D., Shi, H., Clausse, A., Naumann, M., Kümmerle, J., Königshof, H., Stiller, C., Fortelle, A. d. L., and Tomizuka, M. (2019). INTERACTION dataset: An INTERnational, adversarial and cooperative moTION dataset in interactive driving scenarios with semantic maps. *arXiv preprint:1910.03088*.