***********: the central point of contact for data sets in the field of autonomous driving

Anonymous Author(s)
Submission Id: 200

ABSTRACT

The importance of autonomous driving, one of the largest topics within the robotics domain, is well known among researchers as well as the industry. Large amounts of money are invested in research into autonomous driving every year. Associated with this, domain knowledge is also steadily increasing. As a consequence thereof, the need for data sets is growing enormously. Be it for the development of algorithms or their testing, data sets are applicable in numerous use cases. 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 such data sets and their features. In this paper, we establish a first version of ***********, an online tool that provides an overview of more than 150 existing data sets related to the research field of autonomous driving. The tool enables users to sort and filter the data sets according to currently 15 different categories. *********** remains in constant development so that the content stays up-to-date.

KEYWORDS

autonomous driving, data set, overview, collection

ACM Reference Format:
Anonymous Author(s). 2022. ***********: the central point of contact for data sets in the field of autonomous driving. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, New Zealand, May 9–13, 2022*, IFAAMAS, 9 pages.

1 INTRODUCTION

Be it for Mary Barra "My goal is for General Motors to lead in safe autonomous driving" [76] or Elon Musk "Self-driving cars are the natural extension of active safety and obviously something we think we should do" [82] or the ever coming news on further partnerships and investments [54][99][91][45], the importance of autonomous driving technology is very clear to both industry and science.

One of the building stunts 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 suitable, fast and cost-efficient tool on they way to achieving the goal of fully autonomous vehicles.

Against the background of 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 missing relevant data sets. Alternatively they come in the form of scientific papers, therefore slowly but steadily becoming outdated, 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.

In this publication we present our attempt to address this unanswered challenge. With ***********[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], *********** comprises 151 data sets.

Furthermore, *********** 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. ***********, hosted via Github Pages, enables this particularly easily via pull requests.

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

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.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Auckland, New Zealand.

[1] https://***********.com/
[2] as of Oct 07, 2021

Dataset Overview	Entries	Filterable	Sortable	Community Contribution	Number of Itemized Properties	Last Update
***********	**151**	**Yes**	**Yes**	**Yes**	**15**	**2021**
RList	9	**Yes**	**Yes**	No	5	**2021**
Scale	50	**Yes**	No	No	4	2019*
Dataset list	25	**Yes**	No	**Yes**	4	**2021**
YonoStore	10	No	**Yes**	No	-	**2021**
Bifrost	50	**Yes**	**Yes**	No	-	2020*
Kaggle	31	**Yes**	**Yes**	**Yes**	-	**2021**

Table 1: Comparison between ********* 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.**
*Date of publication of the latest entry, as no explicit information was given about the last update

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 the publication by Yin and Berger [103] from 2017, which presents 27 data sets prevalent at the time. Around three times that amount was summarized in a work by Laflamme et al. [75] in 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. In 2020, Feng et al. [62] 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 multi-modality 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. In 2021, another work by Heidecker et al. [68] was published, which 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, a paper by Kim and Hwang [73] 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 their 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 a scientific work. 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 blog [51][93][81], wiki [35][23], git [69][56] or miscellaneous entries [57][74][63], 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 [34] 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* [88] 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* [87] 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 too small to be able to claim a complete overview of existing data sets.

Both *Bifrost* (50 data sets) [43] and *Kaggle* (31 data sets) [32] and especially *YonoStore* (ten data sets) [104] cannot be rated 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* [66], *DeepAI Datasets* [55] or *Papers with Code Datasets* [84] 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 APPROACH

*********** is an online tool designed as a central point of contact for searches for data sets in the field of autonomous driving. It includes a detailed representation of the data sets according to 15 different property categories and enables users to interact via filter and sorting functions.
At the time of publication of this paper, the *********** tool comprises 151 data sets, 40 of which have been examined in detail according to the 15 property categories.

3.1 Content and Structure

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. *********** 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 *********** are those data sets which were published after the famous KITTI data set, which serves, so to speak, as a time benchmark.

[3]Following the definitions of Ulbrich et al. [96]

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 exact breakdown can be seen in table 2.

The selected property categories are in turn the result of an expert survey at the [anonymized] . Over 20 different categories were suggested in the survey (Table 3). Ultimately, 15 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 15 categories are presented in detail at this point.

3.1.1 Size [h]. The category Size [h] describes the size of the data set in hours.

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

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

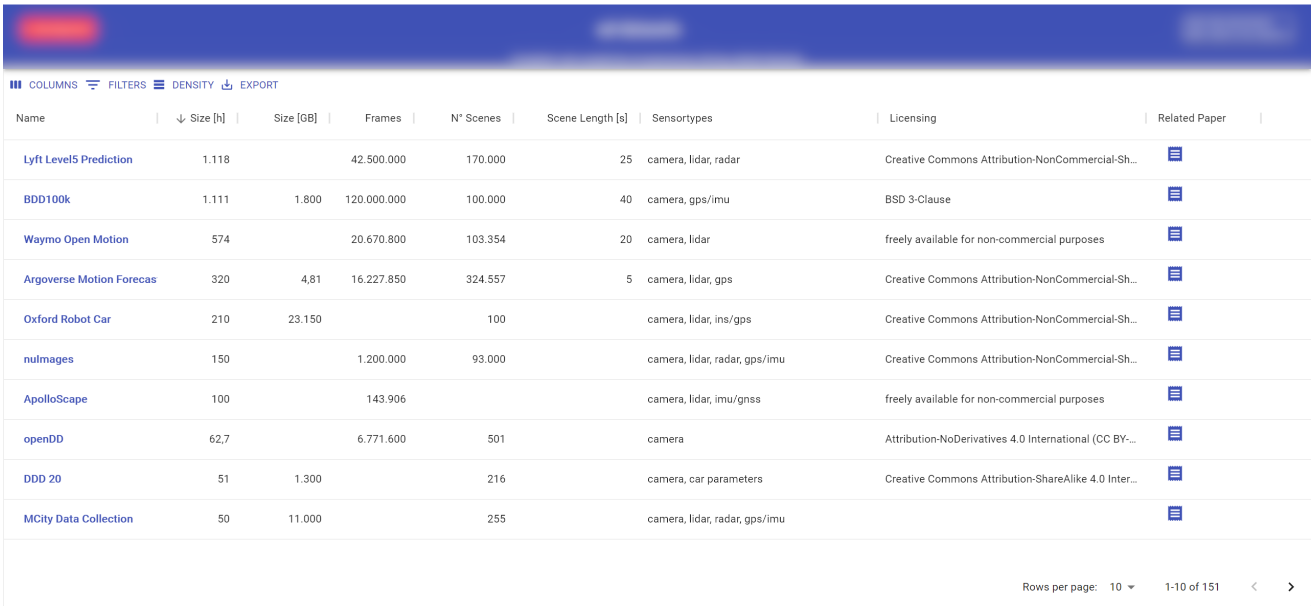
3.1.4 N° Scenes. N° Scenes shows the number of scenes contained in the data set and includes the training, testing and validation segments. In the case of video recordings, one recording corresponds to one scene. For data sets consisting of photos, a photo is the equivalent to a scene.

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

3.1.6 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.

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

3.1.8 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.



Name	Size [h]	Size [GB]	Frames	N° Scenes	Scene Length [s]	Sensortypes	Licensing	Related Paper
Lyft Level5 Prediction	1.118		42.500.000	170.000	25	camera, lidar, radar	Creative Commons Attribution-NonCommercial-Sh...	
BDD100k	1.111	1.800	120.000.000	100.000	40	camera, gps/imu	BSD 3-Clause	
Waymo Open Motion	574		20.670.800	103.354	20	camera, lidar	freely available for non-commercial purposes	
Argoverse Motion Forecas	320	4,81	16.227.850	324.557	5	camera, lidar, gps	Creative Commons Attribution-NonCommercial-Sh...	
Oxford Robot Car	210	23.150		100		camera, lidar, ins/gps	Creative Commons Attribution-NonCommercial-Sh...	
nuImages	150		1.200.000	93.000		camera, lidar, radar, gps/imu	Creative Commons Attribution-NonCommercial-Sh...	
ApolloScape	100		143.906			camera, lidar, imu/gnss	freely available for non-commercial purposes	
openDD	62,7		6.771.600	501		camera	Attribution-NoDerivatives 4.0 International (CC BY-...	
DDD 20	51	1.300		216		camera, car parameters	Creative Commons Attribution-ShareAlike 4.0 Inter...	
MCity Data Collection	50	11.000		255		camera, lidar, radar, gps/imu		

Figure 1: Screenshot of the ********* website (anonymized)**

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

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

3.1.11 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.

3.1.12 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.

3.1.13 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.

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

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

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.

3.2 Technical Implementation

*********** is hosted on Github Pages. This allows for a seamless integration of the research community. Not only can the community influence the development process of the tool, 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 easily be added via pull requests, which automatically triggers the build system to publish the changes on the website.

The implementation of the tool itself was done using the frameworks React [61] and Material-UI [78]. The latter enables a quick and uncomplicated creation of filterable as well as sortable tables. In this work, the data grid component [79] of the Material-UI framework was used.

4 EVALUATION

This section takes a closer look at the 40 data sets from *********** 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 was not chosen randomly, a clear tendency can be identified that both the amount of data sets as well as 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

Name	Size [h]	Frames	N° Scenes	Sensortypes	Publish Date	Sources
Cityscapes 3D[i]				camera, gps, thermometer	2016.02	[53][67][3]
ApolloScape[i]	100	143,906		camera, lidar, imu/gnss	2018.03	[98][8]
Lyft Level5 Prediction[i]	1,118	42,500,000	170,000	camera, lidar, radar	2020.06	[70][29]
Lyft Level5 Perception[i]	2.5		366	camera, lidar	2019.07	[70][28]
Oxford Robot Car[i]	210		100	camera, lidar, ins/gps	2016.11	[77][6]
nuScenes[i]	15	1,400,000	1,000	camera, lidar, radar, gps/imu	2019.03	[47][18]
nuImages[i]	150	1,200,000	93,000	camera, lidar, radar, gps/imu	2020.07	[17]
PandaSet[i]	0.23	48,000	103	camera, lidar, gps/imu	2020.04	[30]
Waymo Open Motion[i]	574	20,670,800	103,354	camera, lidar	2021.03	[60][38]
Waymo Open Perception[i]	10.83	390,000	1,950	camera, lidar	2019.08	[94][21]
KITTI[i]	6		50	camera, lidar, gps/imu	2012.03	[64][2]
BDD100k[i]	1,111	120,000,000	100,000	camera, gps/imu	2020.04	[105][9]
openDD[i]	62.7	6,771,600	501	camera		[46][16]
WildDash[i]			156	camera	2018.02	[106][31]
RoadAnomaly21[i]		100	100	camera	2021.04	[48][37]
Comma2k19[i]	33.65		2,019	camera, radar, gnss/imu	2018.12	[89][52]
KITTI-360[i]		400,000		camera, lidar, gps/imu	2015.11	[102][33]
Fishyscapes[i]				camera	2019.09	[44][20]
LostAndFound[i]		21,040	112	camera	2016.09	[86][4]
Semantic KITTI[i]		43,552	21	lidar	2019.07	[41][19]
KAIST Multi-Spectral Day/Night[i]				camera, lidar, gps/imu, therm. camera	2017.12	[50][7]
A2D2[i]		433,833	3	camera, lidar, gps/imu	2020.04	[65][25]
Caltech Pedestrian[i]	10	1,000,000	137	camera	2010.03	[58][1]
Argoverse Motion Forecasting[i]	320		324,557	camera, lidar, gps	2019.06	[49][13]
Argoverse 3D Tracking[i]			113	camera, lidar, gps	2019.06	[49][13]
KAIST Urban[i]			18	camera, lidar, gps/imu	2017.09	[72][15]
Udacity[i]	10			camera, lidar, gps/imu	2016.09	[95]
Ford Autonomous Vehicle[i]				camera, lidar, gps/imu	2020.03	[39][26]
INTERACTION[i]	16.5	594,588		camera	2019.09	[107][14]
MCity Data Collection[i]	50		255	camera, lidar, radar, gps/imu	2019.12	[59]
Oxford Radar Robot Car[i]			32	camera, lidar, radar, gps/imu	2020.02	[40][83]
NightOwls[j]	5.17	279,000	40	camera	2018.12	[80][11]
DDD 20[j]	51		216	camera, car parameters	2020.02	[71][24]
H3D[j]	0.77	27,721	160	camera, lidar, gps/imu	2019.03	[85]
4Seasons[j]			30	camera, imu/rtk-gnss	2020.10	[100][22]
RadarScenes[j]	4		158	camera, radar, odometry	2021.03	[90][36]
India Driving[j]		10,004	182	camera	2018.11	[97][10]
Synscapes[j]		25,000	25,000	camera	2018.10	[101][12]
RADIATE[j]	5			camera, lidar, radar, gps/imu	2020.10	[92][27]
Bosch Small Traffic Lights[j]		13,427		camera	2017.05	[42][5]

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

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 (Table 4), 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

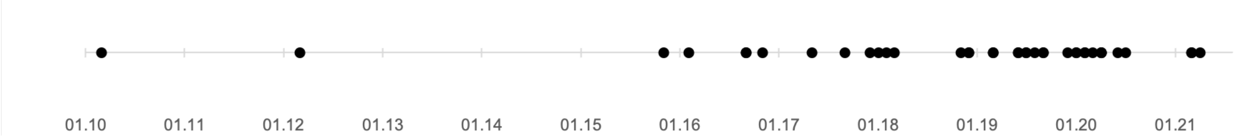


Figure 2: Timeline when data sets have been published

Property	Included	
Size [GB]	Yes	
Size [h]	Yes	
Frames	Yes	
N° Scenes	Yes	
Sampling Rate [Hz]	Yes	
Scene Length [s]	Yes	
Sensor Types	Yes	
Sensors - Details	Yes	
Benchmark	Yes	
Annotations	Yes	
Licensing	Yes	
Related Datasets	Yes	
Publish Date	Yes	
Last Update	Yes	
Related Paper	Yes	
Statistics	No	Future Work
Data Format	No	Future Work
Tooling	No	Future Work
Recording Perspective	No	Future Work
Location	No	Future Work
Main Focus	No	Future Work

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

Camera	Lidar	Radar	Thermometer	Thermal Camera
39	23	8	1	1

Table 4: Overview of use of sensor types in data sets

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 4) 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 3), 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. Thus, 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 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.

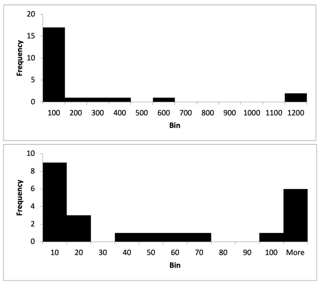


Figure 3: Histogram depicting the distribution of the data sets over their scope of time

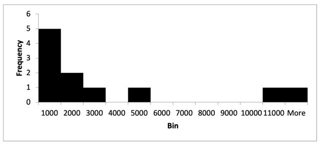


Figure 4: Histogram depicting the distribution of the data sets over their storage size

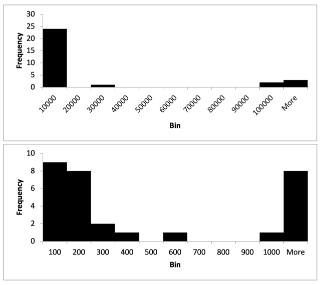


Figure 5: Histogram depicting the distribution of the data sets over their number of scenes

These impressions can also be transferred to the analysis of the number of scenes in data sets (Table 5). 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.

In summary, the majority of the data sets are similar in size. Yet, there is a huge gap in relation to the large data sets.

5 CONCLUSION AND OUTLOOK

With ***********, 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. *********** offers users an overview of over 150 data sets, which are further broken down into 15 different properties. Finally, users can interact with the tool using filter and sorting functions. The up-to-dateness of *********** is maintained through further maintenance of the tool and can be supported by the community, e.g. via pull requests.

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 ongoing work in progress. In addition to the 15 property categories that are already included in the tool, there are further categories that are worth being considered. For one thing, the tool currently does not provide information on the key aspects authors attempt to address with the release of a data set. As research progresses, the relevance of this information increases as well, since the addressed key aspects become more and more specific. For another thing, *********** could be expanded by information on the recording locations of the data sets. Further suggestions include the provision of information on properties such as the data format, the tooling options and the perspective of recording. Statistical evaluations could be included as well, e.g. information on the segmentation of training, testing and validation data.

Finally, it is crucial to obtain feedback from the research community. After all, they are the target audience of ***********. 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 ***********. Therefore, they are subject to future work. It was proposed to include information on the statistical distribution on classes, labels etc. For the time being, this propose remains subject to future work, as there were concerns regarding copyright. Likewise, the suggested property categories data format, tooling options and the perspective of recording 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.

REFERENCES

[1] 2009. Caltech Pedestrian Detection Benchmark. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/
[2] 2012. The KITTI Vision Benchmark Suite. http://www.cvlibs.net/datasets/kitti/
[3] 2016. Cityscapes Dataset – Semantic Understanding of Urban Street Scenes. https://www.cityscapes-dataset.com/
[4] 2016. LostAndFoundDataset - 6D-Vision. http://www.6d-vision.com/lostandfounddataset
[5] 2017. Bosch Small Traffic Lights Dataset. https://hci.iwr.uni-heidelberg.de/content/bosch-small-traffic-lights-dataset
[6] 2017. Oxford Robotcar Dataset. https://robotcar-dataset.robots.ox.ac.uk/
[7] 2017. Visual Perception for Autonomous Driving. https://sites.google.com/view/multispectral
[8] 2018. Apollo Scape. http://apolloscape.auto/
[9] 2018. BDD100K. https://www.bdd100k.com/
[10] 2018. India Driving Dataset. https://idd.insaan.iiit.ac.in/
[11] 2018. NightOwls dataset – Pedestrians at night. https://www.nightowls-dataset.org/
[12] 2018. SynScapes. https://7dlabs.com/synscapes-overview
[13] 2019. Argoverse BY ARGO AI. https://www.argoverse.org/
[14] 2019. Interaction Dataset. https://interaction-dataset.com/
[15] 2019. KAIST Urban. https://irap.kaist.ac.kr/dataset/
[16] 2019. L3Pilot: OpenDD. https://l3pilot.eu/data/opendd
[17] 2019. nuImages. https://www.nuscenes.org/nuimages
[18] 2019. nuScenes. https://www.nuscenes.org/

[19] 2019. SemanticKITTI - A Dataset for LiDAR-based Semantic Scene Understanding. http://www.semantic-kitti.org/
[20] 2019. The Fishyscapes Benchmark - The Fishyscapes Benchmark. https://fishyscapes.com/
[21] 2019. Waymo Open Dataset Perception. https://waymo.com/open/data/perception/
[22] 2020. 4Seasons: A Cross-Season Dataset for Multi-Weather SLAM in Autonomous Driving . https://www.4seasons-dataset.com/
[23] 2020. Automated Driving Datasets - FOT-Net WIKI. https://wiki.fot-net.eu/index.php/Automated_Driving_Datasets
[24] 2020. DDD20: end-to-end DAVIS driving dataset. https://sites.google.com/view/davis-driving-dataset-2020/home
[25] 2020. Driving Dataset | a2d2.audi. https://www.a2d2.audi/a2d2/en.html
[26] 2020. Ford Autonomous Vehicle Dataset. https://avdata.ford.com/
[27] 2020. Heriot-Watt RADIATE Dataset. http://pro.hw.ac.uk/radiate/
[28] 2020. Lyft Level 5 Open Data Perception Dataset. https://level-5.global/data/perception/
[29] 2020. Lyft Level 5 Open Data Prediction Dataset. https://level-5.global/data/prediction/
[30] 2020. PandaSet - Open-source Dataset for Self-driving Cars. https://pandaset.org/
[31] 2020. WildDash 2 Benchmark. https://wilddash.cc/
[32] 2021. Datasets. https://www.kaggle.com/datasets?search=Autonomous+Driving&sort=updated
[33] 2021. KITTI-360. http://www.cvlibs.net/datasets/kitti-360/
[34] 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
[35] 2021. List of datasets for machine-learning research. https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research#Object_detection_and_recognition
[36] 2021. Radar Scenes. https://radar-scenes.com/
[37] 2021. Segment Me If You Can - Datasets. https://segmentmeifyoucan.com/datasets
[38] 2021. Waymo Open Dataset Motion. https://waymo.com/open/data/motion/
[39] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James McBride. 2020. Ford Multi-AV Seasonal Dataset. _eprint: 2003.07969.
[40] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. 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)* (Paris). https://arxiv.org/abs/1909.01300
[41] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*.
[42] Karsten Behrendt and Libor Novak. 2017. A Deep Learning Approach to Traffic Lights: Detection, Tracking, and Classification. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE.
[43] Bifrost. 2020. Search for visual datasets. https://datasets.bifrost.ai/
[44] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Siegwart, and Cesar Cadena. 2019. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *CoRR* abs/1904.03215 (2019). arXiv:1904.03215 http://arxiv.org/abs/1904.03215
[45] BMW Group. 2016. BMW Group, Intel and Mobileye Team Up to Bring Fully Autonomous Driving to Streets by 2021. https://www.press.bmwgroup.com/global/article/detail/T0261586EN/bmw-group-intel-and-mobileye-team-up-to-bring-fully-autonomous-driving-to-streets-by-2021
[46] Antonia Breuer, Jan-Aike Termöhlen, Silviu Homoceanu, and Tim Fingscheidt. 2020. openDD: A Large-Scale Roundabout Drone Dataset. *CoRR* abs/2007.08463 (2020). arXiv:2007.08463 https://arxiv.org/abs/2007.08463
[47] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2019. nuScenes: A multimodal dataset for autonomous driving. (2019).
[48] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. 2021. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. abs/2104.14812 (2021). https://arxiv.org/abs/2104.14812 _eprint: 2104.14812.
[49] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. 2019. Argoverse: 3D Tracking and Forecasting with Rich Maps. abs/1911.02620 (2019). http://arxiv.org/abs/1911.02620 _eprint: 1911.02620.
[50] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. 2018. KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving. *IEEE Transactions on Intelligent Transportation Systems* 19, 3 (2018), 934–948. https://doi.org/10.1109/TITS.2018.2791533
[51] Ambika Choudhury. 2020. Top 10 Popular Datasets For Autonomous Driving Projects. https://analyticsindiamag.com/top-10-popular-datasets-for-autonomous-driving-projects/
[52] comma.ai. 2019. GitHub - commaai/comma2k19: A driving dataset for the development and validation of fused pose estimators and mapping algorithms. https://github.com/commaai/comma2k19
[53] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. abs/1604.01685 (2016). http://arxiv.org/abs/1604.01685 _eprint: 1604.01685.
[54] Daimler. 2020. Mercedes-Benz and NVIDIA:. https://www.daimler.com/innovation/product-innovation/autonomous-driving/mercedes-benz-and-nvidia-plan-cooperation.html
[55] DeepAI. 2017. DeepAI - Datasets. https://deepai.org/datasets
[56] Manfreddiaz Diaz. 2021. GitHub - manfreddiaz/awesome-autonomous-vehicles. https://github.com/manfreddiaz/awesome-autonomous-vehicles
[57] DIY Robocars. 2017. Open Datasets. https://diyrobocars.com/open-datasets/
[58] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 304–311.
[59] Yiqun Dong, Yuanxin Zhong, Wenbo Yu, Minghan Zhu, Pingping Lu, Yeyang Fang, Jiajun Hong, and Huei Peng. 2019. Mcity Data Collection for Automated Vehicles Study. abs/1912.06258 (2019). http://arxiv.org/abs/1912.06258 _eprint: 1912.06258.
[60] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Benjamin Sapp, Charles Qi, Yin Zhou, Zoey Yang, Aurelien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. 2021. Large Scale Interactive Motion Forecasting for Autonomous Driving : The Waymo Open Motion Dataset. abs/2104.10133 (2021). https://arxiv.org/abs/2104.10133 _eprint: 2104.10133.
[61] Facebook. 2013. React – A JavaScript library for building user interfaces. https://reactjs.org/
[62] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems* 22, 3 (2020), 1341–1360.
[63] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. 2021. Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. https://boschresearch.github.io/multimodalperception/dataset.html
[64] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets Robotics: The KITTI Dataset. (2013).
[65] 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. 2020. A2D2: Audi Autonomous Driving Dataset. (2020). https://www.a2d2.audi _eprint: 2004.06320.
[66] Google. 2018. Dataset Search. https://datasetsearch.research.google.com/search?query=Autonomous%20driving&docid=L2cvMTFwd2Y0amZ0Yw%3D%3D
[67] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. 2020. Cityscapes 3D: Dataset and Benchmark for 9 DoF Vehicle Detection. abs/2006.07864 (2020). https://arxiv.org/abs/2006.07864 _eprint: 2006.07864.
[68] Florian Heidecker, Jasmin Breitenstein, Kevin Rösch, Jonas Löhdefink, Maarten Bieshaar, Christoph Stiller, Tim Fingscheidt, and Bernhard Sick. 2021. An Application-Driven Conceptualization of Corner Cases for Perception in Highly Automated Driving. *arXiv preprint arXiv:2103.03678* (2021).
[69] Lhyfst Heyuan. 2019. GitHub - lhyfst/awesome-autonomous-driving-datasets. https://github.com/lhyfst/awesome-autonomous-driving-datasets
[70] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. 2020. One Thousand and One Hours: Self-driving Motion Prediction Dataset. abs/2006.14480 (2020). https://arxiv.org/abs/2006.14480 _eprint: 2006.14480.
[71] Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbrück. 2020. DDD20 End-to-End Event Camera Driving Dataset: Fusing Frames and Events with Deep Learning for Improved Steering Prediction. abs/2005.08605 (2020). https://arxiv.org/abs/2005.08605 _eprint: 2005.08605.
[72] Jinyong Jeong, Younggun Cho, Young-Sik Shin, Hyunchul Roh, and Ayoung Kim. 2019. Complex urban dataset with multi-level sensors from highly diverse urban environments. *The International Journal of Robotics Research* 38, 6 (2019), 642–657.
[73] Seong-heum Kim and Youngbae Hwang. 2021. A Survey on Deep Learning Based Methods and Datasets for Monocular 3D Object Detection. *Electronics* 10, 4 (2021), 517.

[74] Krunal Krunal. 2018. Semantic Segmentation Datasets for Urban Driving Scenes. https://autonomous-driving.org/2018/07/15/semantic-segmentation-datasets-for-urban-driving-scenes/
[75] Charles-Éric Noël Laflamme, François Pomerleau, and Philippe Giguere. 2019. Driving datasets literature review. *arXiv preprint arXiv:1910.11968* (2019).
[76] Belinda Luscombe. 2016. 13 Questions with Mary Barra. *TIME* Vol 187 No 22 (June 2016). https://time.com/4354740/mary-barra/
[77] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 2017. 1 Year, 1000km: The Oxford RobotCar Dataset. 36, 1 (2017), 3–15. https://doi.org/10.1177/0278364916679498
[78] Material-UI. 2014. Material-UI: A popular React UI framework. https://material-ui.com/
[79] Material-UI. 2020. React Data Grid component - Material-UI. https://material-ui.com/components/data-grid/
[80] Lukáš Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, and Bernt Schiele. 2018. NightOwls: A pedestrians at night dataset. In *Asian Conference on Computer Vision*. Springer, 691–705.
[81] T. Nguyen's. 2021. List of public large-scale datasets for autonomous driving research. https://tin.ng/public-datasets-for-autonomous-driving-research/
[82] Alan Ohnsman. 2013. Tesla CEO Talking With Google About 'Autopilot' Systems. *Bloomberg* (May 2013). https://www.bloomberg.com/news/articles/2013-05-07/tesla-ceo-talking-with-google-about-autopilot-systems
[83] Oxford Robotics Institute. 2020. The Oxford Radar RobotCar Dataset. https://oxford-robotics-institute.github.io/radar-robotcar-dataset/
[84] Papers with Code. 2018. Papers with Code - Datasets. https://paperswithcode.com/datasets
[85] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. 2019. The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes. abs/1903.01568 (2019). http://arxiv.org/abs/1903.01568 _eprint: 1903.01568.
[86] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. 2016. Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles. abs/1609.04653 (2016). http://arxiv.org/abs/1609.04653 _eprint: 1609.04653.
[87] N. Plesa. 2021. Dataset list - A list of datasets for machine learning. https://www.datasetlist.com/
[88] Scale. 2019. Open Datasets - Scale. https://scale.com/open-datasets
[89] Harald Schafer, Eder Santana, Andrew Haden, and Riccardo Biasini. 2018. A Commute in Data: The comma2k19 Dataset. _eprint: arXiv:1812.05752.
[90] Ole Schumann, Markus Hahn, Nicolas Scheiner, Fabio Weishaupt, Julius F. Tilly, Jürgen Dickmann, and Christian Wöhler. 2021. RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications. abs/2104.02493 (2021). https://arxiv.org/abs/2104.02493 _eprint: 2104.02493.
[91] Danny Shapiro and NVIDIA. 2019. Mercedes-Benz, NVIDIA to Partner on Next-Gen Vehicles - NVIDIA Blog. https://blogs.nvidia.com/blog/2019/01/08/mercedes-benz-nvidia-next-gen-vehicles/
[92] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. 2020. RADIATE: A Radar Dataset for Automotive Perception. (2020).
[93] Cambridge Spark. 2018. 50 Free Machine Learning Datasets: Self-Driving Cars. https://blog.cambridgespark.com/50-free-machine-learning-datasets-self-driving-cars-d37be5a96b28
[94] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2019. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. abs/1912.04838 (2019). http://arxiv.org/abs/1912.04838 _eprint: 1912.04838.
[95] Udacity. 2020. GitHub - udacity/self-driving-car: The Udacity open source self-driving car project. https://github.com/udacity/self-driving-car/
[96] Simon Ulbrich, Till Menzel, Andreas Reschka, Fabian Schuldt, and Markus Maurer. 2015. Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. 982–988. https://doi.org/10.1109/ITSC.2015.164
[97] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. 2019. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1743–1751.
[98] Peng Wang, Xinyu Huang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. 2019. The apolloscape open dataset for autonomous driving and its application. (2019). Publisher: IEEE.
[99] Michael Wayland. 2020. Volkswagen closes $2.6 billion investment in self-driving start-up Argo AI. (06 2020). https://www.cnbc.com/2020/06/02/vw-closes-2point6-billion-investment-in-self-driving-startup-argo-ai.html
[100] Patrick Wenzel, Rui Wang, Nan Yang, Qing Cheng, Qadeer Khan, Lukas von Stumberg, Niclas Zeller, and Daniel Cremers. 2020. 4Seasons: A Cross-Season Dataset for Multi-Weather SLAM in Autonomous Driving. abs/2009.06364 (2020). https://arxiv.org/abs/2009.06364 _eprint: 2009.06364.
[101] Magnus Wrenninge and Jonas Unger. 2018. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. abs/1810.08705 (2018). http://arxiv.org/abs/1810.08705 _eprint: 1810.08705.
[102] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger. 2016. Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
[103] Hang Yin and Christian Berger. 2017. When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. 1–8. https://doi.org/10.1109/ITSC.2017.8317828
[104] YonoStore. 2021. Datasets. https://store.yonohub.com/product-category/datasets/?filters=product_cat[19]
[105] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. 2018. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. abs/1805.04687 (2018). http://arxiv.org/abs/1805.04687 _eprint: 1805.04687.
[106] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. 2018-09. WildDash - Creating Hazard-Aware Benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
[107] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kümmerle, Hendrik Königshof, Christoph Stiller, Arnaud de La Fortelle, and Masayoshi Tomizuka. 2019. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. abs/1910.03088 (2019). http://arxiv.org/abs/1910.03088 _eprint: 1910.03088.