

Challenges in Designing Datasets and Validation for Autonomous Driving

Michal Uříčář¹^a, David Hurych¹, Pavel Křížek¹ and Senthil Yogamani²^b

¹Valeo R&D DVS, Prague, Czech Republic

²Valeo Vision Systems, Tuam, Ireland

Keywords: Visual Perception, Design of Datasets, Validation Scheme, Automated Driving.

Abstract: Autonomous driving is getting a lot of attention in the last decade and will be the hot topic at least until the first successful certification of a car with Level 5 autonomy (International, 2017). There are many public datasets in the academic community. However, they are far away from what a robust industrial production system needs. There is a large gap between academic and industrial setting and a substantial way from a research prototype, built on public datasets, to a deployable solution which is a challenging task. In this paper, we focus on bad practices that often happen in the autonomous driving from an industrial deployment perspective. Data design deserves at least the same amount of attention as the model design. There is very little attention paid to these issues in the scientific community, and we hope this paper encourages better formalization of dataset design. More specifically, we focus on the datasets design and validation scheme for autonomous driving, where we would like to highlight the common problems, wrong assumptions, and steps towards avoiding them, as well as some open problems.

1 INTRODUCTION

We have the privilege to live in the exciting era of high pace research and development aiming for the full autonomy in the ground transportation, involving all major automotive industries. Nowadays, the sort of standard is the autonomy level 2 (International, 2017). We can see the progress towards levels 3 and 4, and the ultimate goal is, of course, to achieve level 5, i.e., the real full autonomy. In Figure 1, we outline all the levels of autonomy in automotive for reference, as described by (International, 2017).

Naturally, there is a high motivation and willingness to speed up the progress in combination with the recent success of deep neural networks. However, this leads to the development of certain bad practices, which are progressively more and more visible in research papers. The goal of this paper is to determine some of the bad practices, especially those related to the issue of dataset design and validation scheme, and propose the ideas for fixing them. Apart from that, we would also like to identify several open problems for which the standardized solution is yet to be discovered.

The importance of dataset design is often over-

looked in the computer vision community; the problem was addressed in detail in the ECCV workshop in 2016 (Goesele et al., 2016). In (Khosla et al., 2012) the authors discuss issues with dataset bias and how to address them. In general, we can say that having a good and representative data is the crucial problem of virtually all machine learning techniques. Often, the applied algorithms come together with a requirement for data to be independent and identically distributed (i.i.d). However, this requirement is frequently broken and not checked for. Either the dataset parts are obtained from different distributions or their independence is questionable. Also the definition of terms *identically* and *independently* depends on the foreseen application.

Frequently, we can also see that researchers blindly follow the results of the evaluation represented by a key performance indicator (KPI) without making the trouble to check for the experiment correctness. With the increasing size of the datasets, the errors in annotations become significant, and the absence of a careful inspection might be dangerous. Especially, when we achieved the state where the improvements in decimals point are considered essential. A lot of models are discarded during the design, because of the lack of systematical analysis and the effort of getting a real insight. However, such a complete and systematical analysis might render some

^a <https://orcid.org/0000-0002-2606-4470>

^b <https://orcid.org/0000-0003-3755-4245>

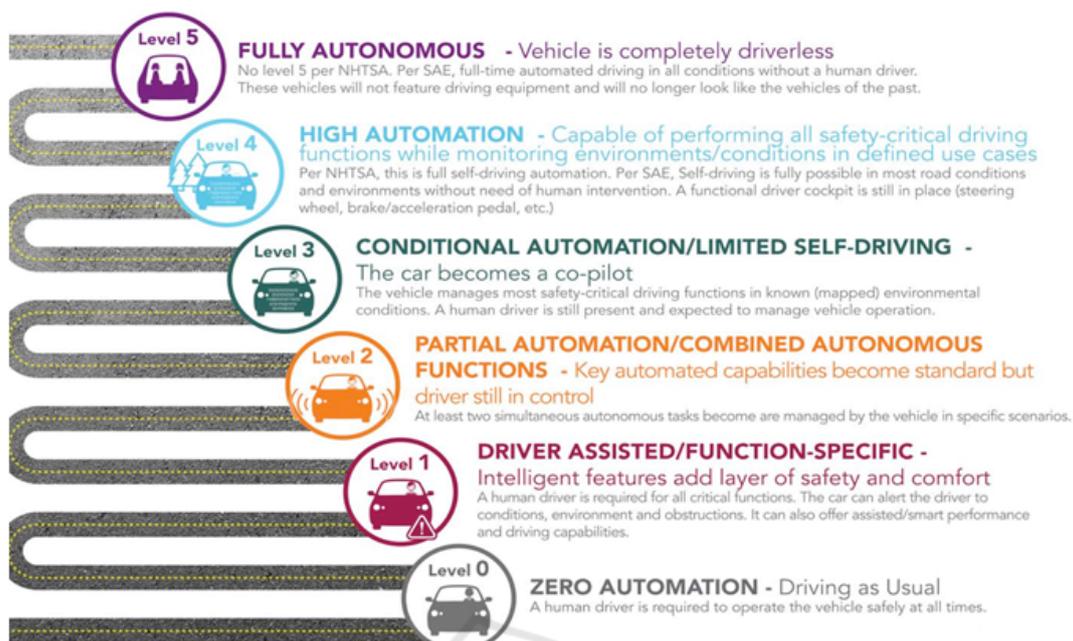


Figure 1: Levels of autonomy as described by (International, 2017). Note, that the standard in 2018 is level 2, since the first sensor allowing level 3 functionality was just released in 2018. We can participate quick adaptation to level 3 in the next year. The ultimate goal is, of course, level 5, for which a lot of companies is aiming now. However, the path to get to the level 5 is still quite long and the predictions are speculating about decades to get there.

of these models more robust or to better generalize across different datasets.

Last but not least, we should emphasize the importance of fair comparisons regarding the used resources or a model complexity. Taking deep neural networks as a gentle example, what if we use an ensemble of simpler classifiers, trying to call the complexity of a neural net? Would we still see the same performance gap? As a typical example of methods and models compared only from the performance point of view while ignoring the computational complexity we may take deep Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012) versus the Deformable Part based Models (DPM) (Felzenszwalb et al., 2010). The CNN models are dominating DPM since their break-through in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) back in 2012. However, we tried to do a bit “fairer” comparison, where the model complexity was more or less matched (i.e. both CNN and DPM models consisted of the similar number of operations in the inference calculation). The results from the independent testing set are summarized in Figure 2, where we experimented with these methods on our internal pedestrian dataset in real driving scenarios. Both CNN and DPM models were trained on the same training set. The dataset splits were obtained by the means of stratified sampling. The DPM model

is better in several different settings.

The rest of the paper is structured as follows. Section 2 discusses the current bad practices in dataset design, emphasizes its importance and lists open issues. Section 3 summarizes issues related to the validation of a safety system. Finally, Section 5 concludes the paper and provides future directions.

2 DATASET DESIGN

Typically in academia, test and validation datasets are provided, and the goal is to get the best accuracy. However, in the industry deployment setting, datasets have to be designed interactively along with the model design. Unfortunately, there is minimal systematic design effort as it is difficult to formulate the problem and quantify the quality of datasets. First, there is a data capture process where cameras mounted on a vehicle capture necessary data. This process has to be repeated across many scenarios like different countries, weather conditions, times of the day, etc.

Table 1 lists the popular automotive datasets for semantic segmentation task. There is visible progress towards the increased number of samples as well as a change in direction to realistic data.

Table 1: Popular automotive datasets for semantic segmentation.

Dataset	CamVid (Brostow et al., 2008)	Citiscapes (Cordts et al., 2016)	Synthia (Ros et al., 2016)
Annotation	700 images	5000 images	200,000 images
Note	Cambridge captures	Germany captures	Synthetic data
Dataset	Virtual KITTI (Gaidon et al., 2016)	Mapillary Vistas (Neuhold et al., 2017)	ApolloScape (Huang et al., 2018)
Annotation	21,260	25,000 images; 100 classes	143,000 images; 50 classes
Note	Synthetic data	Six continents	China captures

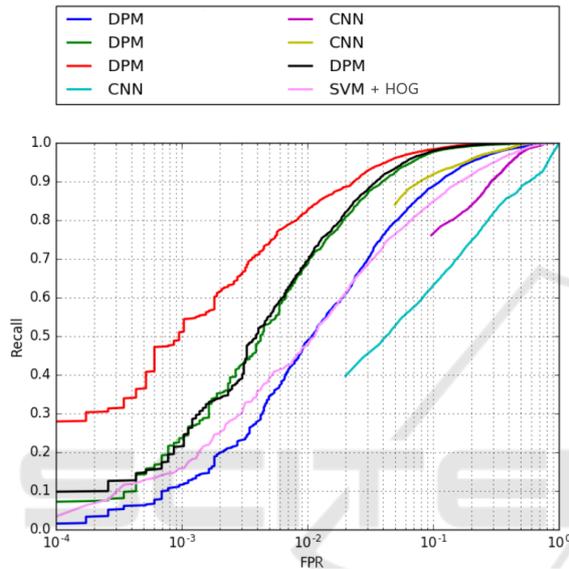


Figure 2: Performance comparison for equally complex CNN and DPM models on a pedestrian detection task. Multiple settings were tested in an attempt to reach best performance for both methods. Due to internal confidentiality rules we can not reveal individual settings. Here only SVM+HOG had lower complexity than other models and served as a baseline. 3 curves do not pass through the whole [Recall, FPR] space because of lack of data-points in next confidence step. We were interested in parts of curves with recall higher than 0.9.

2.1 Typical Scenarios of Dataset Design in Autonomous Driving

There are three scenarios currently possible for automated driving researchers: (i) academic setup where public datasets are used as they are. However, these have commercial licensing restrictions and cannot be used freely for industrial research nor for final production, (ii) proof of concept setup where dataset is collected for a restricted scope, e.g. one city, regular weather conditions, and (iii) a production system where dataset has to be designed for all scenarios, e.g., a large set of countries, weather conditions, etc.

Here, we focus our discussion on the third type.

A typical process for dataset design comprises of the following steps. Firstly, the requirements are created for coverage of countries, weather conditions, object diversity, etc. Then video captures are acquired and their frames arbitrarily sampled without a systematic sampling strategy. Next step is creating the training, validation and test splits from all gathered images. This can be done either randomly, or in a better case by stratified sampling, retaining the class distributions among the splits. Then, after the model is trained, one has to evaluate the KPIs, such as the mean intersection over union (mIoU). Last but not least step is a naïve search for corner cases in the test split and addition of such infected samples to the training split (in best case by obtaining new data, which look similar to the testing ones).

3 VALIDATION

Validation scheme is another critical part of the current research. In the automotive industry, this topic is even more critical, due to the very stringent requirements on safety. However, we would like to emphasize here, that the common problems with validation are shared among other fields as well.

The AD systems have unique criteria due to functional safety and traceability issues. The artificial intelligence software for the AD has to comply with strict processes like ISO26262 to ensure functional safety. Thus, apart from accuracy validation, it is essential to do rigorous testing of software stability. However, unit testing of AI algorithms comes with additional challenges, like large dimensionality of data, or the abstract nature of the model and automated code generation. Due to these challenges, it is difficult to write tests manually. There have been attempts to generate tests automatically using deep learning like DeepXplore (Pei et al., 2017) and DeepTest (Tian et al., 2018) focused on the AD.

3.1 The Need for Virtual Validation and their Limitations

In the automotive industry, one tests the algorithms by recording a lot of hours of various scenarios which should ideally cover all possible real-world situations. Afterward, these hours of recordings are statistically evaluated, and the algorithm is allowed to go for the start of production only if fulfilling the strict requirements, which were formulated and ideally also fixed at the beginning of the project. Despite sounding completely legitimate at first, such an approach has several important flaws. The first, and probably also the most important one is the physical and practical impossibility to cover all real-world scenarios. Let us formulate an example explaining our claims: we want to test the automated parking functionality enhanced by pedestrian detection algorithm. The vehicle should send a command for braking if there is a pedestrian within critical distance and on the collision course present. Now, imagine a legitimate scenario, where the pedestrian should be a toddler sitting right behind the car. In most of the countries across the whole globe, their laws prohibit making of such recordings. The second problem is in setting up the requirements at the beginning of the project. If we agree that it is not possible to record all real-world situations, freezing the requirements tends to influence the scenarios to record as well as their complexity.

One might say, that the solution is obvious—use virtual validation or some other workaround, like usage of dummy objects. However, doing so bring other problems, such as the realism of these artificial scenarios. It is clear that we will never be able to evaluate it for some of the real-world situations.

A common problem (not only in the automotive industry) is the design of the validation scheme itself. Typically, we can see that the algorithm was optimized for a specific criterion, using a particular loss function. Ideally, we should see the same loss function used in the evaluation, alas, it is not uncommon to see something not even similar to the original loss used in testing. The problem with this setup is that it is not possible to optimize for a beforehand unknown criterion.

3.2 Model Survivorship Bias

This mistreated optimization is connected to another significant problem—overfitting to (not only) standard datasets. In research, but also in industry, only the models which are obtaining the best results are reported or survive. Nobody tells, how many times the model failed on those data before it was tweaked

enough to provide the best results. Not reporting the negative results is counter-productive (Borji, 2018). With the increase of the deep learning models in the game, this problem is even worse. Deep networks are known to easily fit random labels, even for randomly generated images (Zhang et al., 2016).

3.3 Complete Reporting and Replicability

Another, and unfortunately also very frequent, problem is that certain important statistics are not being reported. Only rarely, we can see that authors of the research paper give away also their splits of datasets to training/validation/testing parts. Quite frequently, they do not even bother mentioning the key statistics, such as the number of samples they used, or how the splits were obtained. This problem is connected with a choice of machine learning method, which is not well justified. For example, a lot of techniques come with an assumption about the data distribution. Important and integral part of each experimental evaluation should be its replicability by a non-involved party.

3.4 Cross-validation and the Law of Small Numbers

The emphasis on early deployment in the automotive industry often leads to unjustified design choices, which do not have support in data. The infamous law of small numbers gets into the practice. Usually, the researchers and developers have to deal with the insufficient amount of data to do a statistical evaluation. Then, due to the lack of time, some early decisions are made. The smaller the sample of data was used for performing the evaluation, the higher the probability of wrong outcomes. Just imagine you have a fair coin, so the expected probability of getting head after flipping it is 0.5. Let us conduct the following experiment. Flip the coin ten times and count the number of heads/tails. Very likely, you will not get the five heads. Now flip the same coin a thousand times, marking the number of heads. This time, the number of heads will be close to five hundred. If one does only the first experiment, he might be in temptation to question the fairness of the coin. While after the second one, such conduct feels unjustified. It is common to see quite a small number of samples for some particular tasks, as well as only one evaluation over a single split of data.

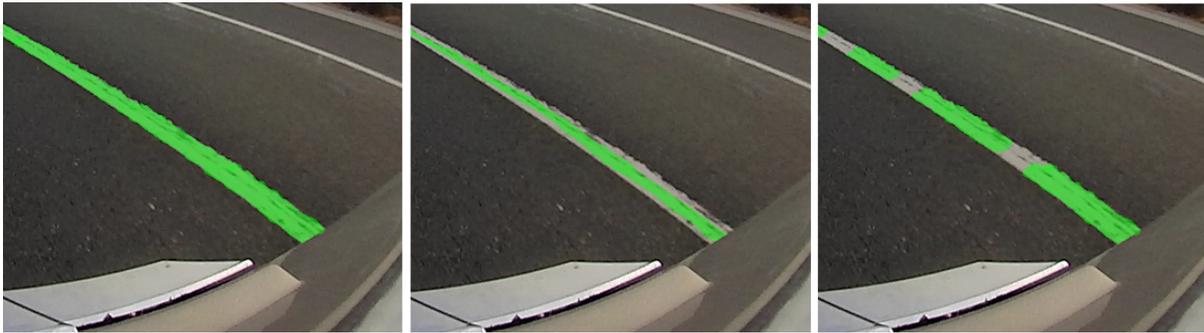


Figure 3: Wrong metrics can lead to misinterpretation of the results. Left: ideal segmentation of the lane marking—whole line is correctly covered. Middle: one possible case of a realistic segmentation obtained from a learned model. Note that the segmented marking is thinner than it should be. However, it is continuous non-interrupted line as it should be. Right: another possible case of a realistic segmentation. This time, the width of the lane marking is correctly segmented. However, there are interruptions, which render the marking looking like a dashed line. Note, that such misinterpretation might be grievous, since as we know both type of markings are admissible having completely different semantic meaning. The example on the rightmost image has higher IoU value, than the middle one.

3.5 The Need for Customized Evaluation Metrics

Standard performance measures, such as mIoU for semantic segmentation, may not translate well to the end user application needs. Let us take the automated parking functionality with pedestrian detection as an example again. A perfect segmentation of a pedestrian is not necessary, and just a coarse detection is sufficient for initiating the braking. Another example is recognition of the lane markings—there are nice examples, where a higher mIoU does not necessarily lead to a better segmentation of the main shape of the marking, which is crucial for its recognition. In Figure 3, we depict one of such examples. In both of these cases, custom tailored evaluation metric is the key for a better algorithm. And visual checks of the results are a must.

4 DISCUSSION

In this section, we would like to discuss several important open issues and suggestions for improvement.

4.1 Open Issues and Suggestions for Improvement

Many visual perception tasks, like semantic segmentation, need a very expensive annotation, leading to unnecessarily smaller datasets. Synthetic datasets, like (Ros et al., 2016), (Gaidon et al., 2016), (Dosovitskiy et al., 2017), (Mueller et al., 2018), can be useful as potential mitigation of the lack of data. However, domain adaptation is usually required, and it is not

clear what ratio of synthetic to real data would be still beneficial.

Due to the popularity of the AD, the available datasets are snowballing, and the choice of datasets starts to be a problem itself. Moreover, in research, there is no synchronization on datasets, and it is difficult to compare different works justly. It might be helpful if the community agrees on some standardized dataset (combining the strengths and weaknesses of all of them) to have a possibility to compare algorithms more thoroughly and honestly. For example, there is no available dataset with wide-angle fisheye camera images. Such a camera is a standard in the AD for capturing the 360° view around the vehicle. A publicly available dataset with multiple cocoon cameras, which are typical for the AD, is also missing.

An automated sampling mechanism for acquiring the training images the goal of which is to get rid of redundant samples and providing a maximal diversity is an open problem. The dataset and model design is done iteratively, and samples are added on the go to improve the model performance. This process is dangerous since such an approach might easily break the i.i.d. requirements on data. A corner case mining is a related topic, where difficult samples are identified, and their knowledge is used for improving the performance. Note, that this process usually takes into the account the sample similarity in the image space. However, one would benefit if the mining would be based on the similarity measured in the classifier's feature space.

Data augmentation is an important mechanism to obtain samples for difficult or rare scenarios. We can take an automated parking system with pedestrian detection as an example. We would like to have data where children are playing with a ball and sometimes

blindly follow the ball which gets on the collision path with the vehicle. One such possible situation is depicted in Figure 4. It is clear that we cannot record such scenario, due to the safety and legal issues. We see the possibility in bypassing such scenario by recording in a controlled environment and applying GANs (Goodfellow et al., 2014) for the domain transfer to fit the AD needs (Chan et al., 2018), (Hoffman et al., 2018).



Figure 4: Typical situation, which is required to be covered by data, but which is also prohibited to arrange for recording by law—a child is playing with a ball, focusing on the play and not paying any attention to the car, which is parking and on the collision path with the child.

5 CONCLUSIONS

In this paper, we attempt to emphasize the importance of dataset design and validation for the AD systems. Both dataset design and validation are highly overlooked topics which have created a large gap between academic research and industrial deployment setting. There is a considerable effort to go from a model which achieves state-of-the-art results in an academic context to the development of a safe and robust system deployed in a commercial car. Unfortunately, there is very little scientific effort spent in this direction. We have tried to summarize the bad practices and listed open research problems based on our experience in this area for more than ten years. Hopefully, this encourages further scientific research in this area and places a seed for future improvement.

REFERENCES

- Borji, A. (2018). Negative results in computer vision: A perspective. *Image Vision Comput.*, 69:1–8.
- Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer.
- Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. (2018). Everybody dance now. *arXiv preprint arXiv:1808.07371*.
- 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 arXiv:1604.01685*.
- Dosovitskiy, A., Ros, G., Codevilla, F., López, A., and Koltun, V. (2017). CARLA: an open urban driving simulator. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, pages 1–16.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.
- Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349.
- Goesele, M., Waechter, M., Honauer, K., and Jaehne, B. (2016). ECCV 2016 Workshop on Datasets and Performance Analysis in Early Vision. online.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial networks. *CoRR*, abs/1406.2661.
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholm, Sweden, Stockholm Sweden. PMLR.
- Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., and Yang, R. (2018). The apolloscape dataset for autonomous driving. *arXiv preprint arXiv:1803.06184*.
- International, S. (2017). The wayback machine.
- Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114.

- Mueller, M., Dosovitskiy, A., Ghanem, B., and Koltun, V. (2018). Driving policy transfer via modularity and abstraction. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, pages 1–15.
- Neuhold, G., Ollmann, T., Bulò, S. R., and Kotschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 5000–5009.
- Pei, K., Cao, Y., Yang, J., and Jana, S. (2017). Deepxplore: Automated whitebox testing of deep learning systems. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18. ACM.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Tian, Y., Pei, K., Jana, S., and Ray, B. (2018). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*, pages 303–314. ACM.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530.

