Towards a Methodology for Training with Synthetic Data on the Example of Pedestrian Detection in a Frame-by-Frame Semantic Segmentation Task

Atanas Poibrenski
German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus
Atanas.Poibrenski@dfki.de

Janis Sprenger
German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus
Janis.Sprenger@dfki.de

Christian Müller
Head of Competence Center for Autonomous Driving, DFKI, Saarland Informatics Campus
Christian.Mueller@dfki.de

ABSTRACT

In order to make highly/fully automated driving safe, synthetic training and validation data will be required, because critical road situations are too divers and too rare. A few studies on using synthetic data have been published, reporting a general increase in accuracy. In this paper, we propose a novel method to gain more in-depth insights in the quality, performance, and influence of synthetic data during training phase in a bounded setting. We demonstrate this method for the example of pedestrian detection in a frame-by-frame semantic segmentation class.

CCS CONCEPTS

• **Computing methodologies** → *Image segmentation*; *Procedural animation*;

KEYWORDS

Semantic Segmentation, Synthetic Data, Automated Driving

ACM Reference Format:
Atanas Poibrenski, Janis Sprenger, and Christian Müller. 2018. Towards a Methodology for Training with Synthetic Data on the Example of Pedestrian Detection in a Frame-by-Frame Semantic Segmentation Task. In *SEFAIAS'18: SEFAIAS'18:IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems , May 28, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3194085.3194093

1 INTRODUCTION

Taking into consideration the manifold activities in this area, it can be regarded as an accepted fact in the community that synthetic data is required in order to make AI for autonomous driving safe. Critical road situations are too rare and too divers – we cannot find enough real data for training or validating, even when collecting another few million miles. Road situations apparently follow a long-tail distribution: relatively few situations happen very often (like "car approaching from the right at an intersection"), while a large number situations happen only rarely (like "child running in front of car"). Also, for many critical situations (like "child running in front of car") we do not even want to collect real data.

Moreover, for training an AI module, it is not sufficient to only have one variant of each situation. Rather than that, all possible variants need to be created, such as "child with green anorak running in front of green car at night in light rain". The amount of data necessary to cover all relevant instances of critical situations both for training as well as validating the AI modules is huge. The number of kilometers needed to collect enough data is estimated to be equivalent to 500 times the distance from Earth to sun and back. [5]

However, in order to use synthetic data, it is necessary to study the characteristics of such data in the context of training and testing of AI modules. In the recent years, a few articles were published that deal with this matter (see section 3). The methodology applied there was mainly adding synthetic data to the training set and tracking overall improvement of the algorithm with respect to a given problem. In this paper, we follow a similar path, taking camera-based frame-by-frame semantic segmentation as our example problem (see section 2). Frame-by-frame semantic segmentation is well defined, there exist common datasets and challenges, benchmarks, and it is a relevant task for object recognition. More specifically, we look at pedestrian detection, because when introducing autonomous driving in city traffic, pedestrians (together with bicyclists, scooter drivers, etc.) belong to the group of most vulnerable road users that need to be protected. Creating synthetic data for pedestrians covers a series of relevant problems including modeling behavior based on intentions, modeling motion based on behavior, and rendering sensor data based on motion.

The main contribution of this paper tackles the problem of finding the right methodology. The approaches in literature lack a notion of relevance of the improvement relative to the effort of creating just some more real training data. In other words, it is not entirely certain that finding an overall improvement justifies the introduction of synthetic data. There are furthermore no details given about how synthetic data "behaves" compared to real data. We introduce the basic foundations of a methodology for experimenting with synthetic training data. We create lower and upper performance "boundaries" [1] and introduce application-motivated analyses of the classification result. Particularly, for the pedestrian detection task, we apply a distance measure and analyze the accuracy depending on the distance to the camera (of the ego vehicle).

SEFAIAS'18, May 28, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5739-5/18/05…$15.00
https://doi.org/10.1145/3194085.3194093

[1] The word "boundaries" in quotation marks because new results can get lower and higher respectively. We could therefore also call it lower and upper reference points.

We believe that this perspective opens up new ways for more detailed analyses of training with synthetic data in the future (see section 6).

2 SEMANTIC SEGMENTATION

Frame-by-frame semantic segmentation is one of the standard tasks in the context of deep learning for environment perception in highly/fully automated driving. Given a single frame of a video source, the task is to determine the semantic target class of each pixel within this frame. The whole image is densely separated into different segments and simultaneously each segment is assigned a semantic class, thus using conceptual differences of objects for the segmentation task. Since semantic segmentation is directly using the sensor data (e.g. RGB-camera) its main task in the context of autonomous driving is the generation of more abstract representations for high level services, for instance object recognition and scene understanding.

There are multiple challenges published, consisting of a training- and an evaluation-set of densely labeled 2D RGB images displaying real street scenes viewed from the perspective of a driving car. The label information assigns each pixel to a semantic class (e.g. Street, Sidewalk, Person, Car, etc.) thus providing enough data to train and evaluate a neural network. In this work we are using the Cityscapes dataset [2] which provides ground-truth depth maps in addition to the semantic labels. This enables us to analyze the results of semantic segmentation in regard to the distance of the object to the camera (ego vehicle).

3 RELATED WORK

Compared to image classification, frame-by-frame semantic segmentation has not received much attention in terms of training and domain adaptation with synthetic data. In terms of machine learning, domain adaptation tries to reduce the mismatch between the real and the synthetic distributions.

The most basic approach is to first train on the source (synthetic) domain and then fine-tune on the real (target) domain as it was done in [10]. An alternative method is to jointly train on both domains by using mini-batches from the source and target domain [7]. This strategy allows for domain-invariant features to be extracted and to bring useful information from both domains resulting in a better performance on the target domain. Another method, used by the creators of the large synthetic dataset Synthia, is to build batches from both domains (real and synthetic) in a fixed ratio where the real images dominate the distribution, while the synthetic ones are used as a sophisticated regularization term [8]. [3] is the first work to address domain adaptation in segmentation models algorithmically and more in-depth. She uses both global and local domain alignment techniques. The most recent work by [12] proposes a curriculum-style learning approach to minimize the domain gap between real and synthetic images. This is done by learning global label distributions over images and local distributions over landmark superpixels.

All of these works focus on the technical aspect of domain adaptation and lack the analysis of the effort of creating/adapting synthetic data versus creating more real data in terms of overall performance.

4 METHOD

In order to analyze the quality and value of synthetic data for a specific test, we claim that it is not sufficient to add a large amount of training data to the original real world images of a real dataset. We propose a method to analyze the synthetic data in a bounded setting. Based on our first assumption that a model trained on real data will – at least for now – outperform models trained on synthetic data, we set the upper bound of the performance measure to the performance of a model trained on 100% of the available training data of the real world dataset. Based on our second assumption, that removing training data should reduces the performance of a model, we train additional models on a fraction $x\%$ of this training data and set the lower bound to the performance of the additional models. Replacing the removed data $(100 - x\%)$ with synthetic images yields models trained on a combination of real and synthetic data, that should show a performance in between the upper and lower bounds. Following this principle, we are keeping the total amount of training data constant and can specifically analyze the benefit and qualities of a single synthetic dataset.

In these experiments we are using intersection over union (IoU, see equation 1) as a performance measure.

$$IoU = \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}} \quad (1)$$

This performance measure was chosen because it takes into account both the false alarms and the missed values of each class, giving us more information of what is happening compared to a simple accuracy measure. Moreover, it has become a common approach for measuring the performance of semantic segmentation algorithms.

5 EXPERIMENTS

We want to analyze the influence of synthetic data on the segmentation performance specifically for pedestrian segmentation, since people are the most versatile aspect of urban driving and are the factor that most likely needs to be simulated to make autonomous driving safe. We have the (straightforward) hypothesis that pedestrians further away from the ego vehicle will be worse classified as pedestrians close to the camera. This deterioration of performance might obscure the difference between real only and combined real and synthetic training data, since we do not yet know in which way the synthetic data works. Thus we performed a first experiment to verify the distance hypothesis and conducted a second experiment on top of the first one to analyze the impact of synthetic data with respect to the distance of pedestrians to the ego vehicle.

5.1 Experiment 1: Pedestrian Segmentation and Distance to the Camera

We trained a dilated VGG16 model [11], that was initialized on the pre-trained weights from ImageNet [9], on the training set of Cityscapes dataset [2] (2975 images). The training was done with the standard mini-batch stochastic gradient descent with momentum. Since the full images are too big to fit into GPU memory, a random crop of 628x628 is used. The mini-batch size is set to 12, the learning rate is 0.0001 and the momentum is 0.99. The training of the model was done using the Caffe framework [4] and a Tesla P100 GPU (16gb).

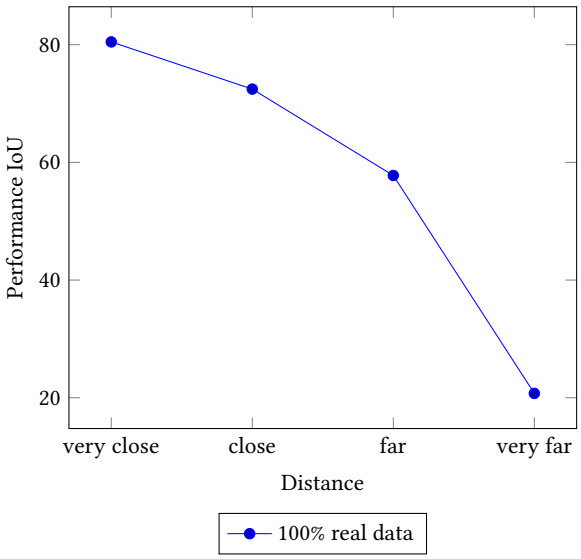

Figure 1: Performance for of Model trained on 100% of the Cityscapes training set for the pedestrian class.

For the evaluation of our model we split all pedestrians inside the evaluation set of Cityscapes into 4 distance groups (very close: $< 5m$, close: $5 - 10m$, far: $10 - 15m$, very far: $> 15m$), based on the disparity maps provided by Cityscapes. The trained model is evaluated on all images (N = 500) and the mean intersection over union for each distance group was calculated. The evaluation result can be seen in figure 1 and show a clear drop in performance, if we increase the distance of pedestrians. Thus our hypothesis is verified and pedestrians are detected worse, the further away they are from the camera.

5.2 Experiment 2: Performance of Synthetic Training Data

In order to analyze synthetic data with respect to segmentation of pedestrians, we generated a dataset of 1497 images (1080p) of a camera path through a synthetic, photo-realistic scene containing 32 different animated avatars. Distance of pedestrians to the camera varied and changed over time, as the camera was driving through the scene, thus there are representatives for all distance groups. All images where generated using the Cycles ray-tracing render engine in Blender [1]. Our hypothesis is, that synthetic data can, at least partially, replace real training data.

We used a random permutation of the synthetically generated training data to pre-train 5 dilated VGG16 models [11] (10%: 297 images, 20%: 595 images, 30%: 892 images, 40%: 1190 images, 50%: 1487 images) that were initialized on the pre-trained weights from ImageNet [9]. All models were fine-tuned on randomly sampled images of the Cityscapes training set, such that each model was trained on a total of 2975 images. In addition we trained 5 models on the same randomly sampled images without the pre-training on synthetic data, producing models trained on 90%, 80%, ... 50% of the total amount of 2975 real images.

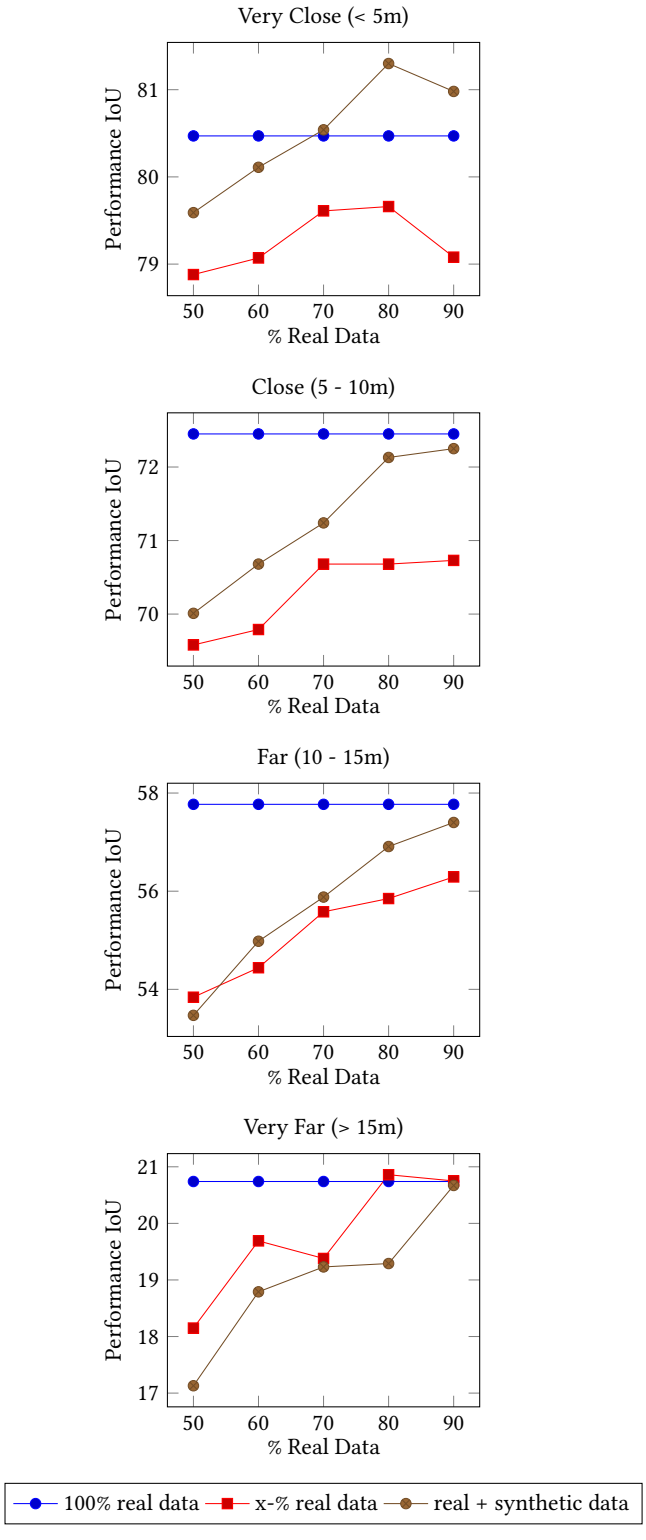

Figure 2: Performance of the model trained on synthetic and real data for the pedestrian class

The configuration of the training algorithm was set to the same values as in experiment 1 and the evaluation scheme was kept the same.

The results are presented in figure 2. The performance in pedestrian segmentation improves in all distance categories except for very far pedestrians. The decline of performance for very far pedestrians can be due to the already poorly performing model or the quality of the training data itself.

6 DISCUSSION

The important observation in Experiments 1 and 2 is not related to the overall accuracy in pedestrian segmentation. It is also not surprising that near pedestrians are better recognized than far away ones. The important observation is the change in performance from the lower boundary towards the upper boundary when replacing real data with synthetic data. The synthetic data seems to "behave" according to our expectations and therefore replacing real training data with synthetic training data will result in a better model compared to leaving the training data out.

In a preliminary experiment we observed the intersection over union aggregated over all pedestrians, which resulted in a best performing model for the pedestrian class trained on 70% of real image data (without synthetic replacements). This contradicted our basic assumption that less training data reduces performance and diminishes the significance of work validating synthetic data only using aggregated performance measures (e.g. [6, 7]), since increasing the training data does not necessarily yield an improvement of performance. The influence of synthetic data can only be observed and explained in a more detailed setting, as presented in this work.

These results should be regarded as a first step, as a test bed for future experiments. We suggest to continue in this direction by systematically modifying the visual quality of the material and studying the behavior. Likewise, it would be an interesting challenge to try creating synthetic training material that helps the (same) network to recognize far away pedestrians and thus specifically tackling a shortcoming of the real training data. In addition we only realized our experiments with a single network structure and although we believe the observed trend should generalize, this remains an open question. Lastly training a neural network includes a certain amount of randomness and hence the result should be validated against statistical certainty.

7 SUMMARY

In this paper, we mainly introduced a methodology for training with synthetic data. We believe that using upper and lower boundaries makes the analysis of differences in the results more informative and allows a better judgment of the "behavior" of synthetic data. In order to explain the impact of synthetic data and the behavior of the network itself it is not sufficient to only consider the aggregated performance measures that are used for comparisons between networks. A set of well designed and bounded experiments however can eminently improve our understanding of the network and the data and thus form premises for future dataset generation.

The method and the results presented build a framework, on-top of which we will explore further relevant visual characteristics for training, analyze similarities between human visual perception and neural networks and gradually move towards the generation of critical situations relevant for the task of fully autonomous driving.

8 ACKNOWLEDGEMENTS

This research was funded in part by the German Federal Ministry of Education and Research under grant number 01/W17003 (project REACT). The responsibility for this publication lies with the authors.

REFERENCES

[1] Blender Online Community. Fri 02/24/2017. *Blender - a 3D modelling and rendering package.* Blender Foundation, Blender Institute, Amsterdam. http://www.blender.org

[2] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3213–3223.

[3] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. 2016. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. *CoRR* abs/1612.02649 (2016). arXiv:1612.02649 http://arxiv.org/abs/1612.02649

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *CoRR* abs/1408.5093 (2014). arXiv:1408.5093 http://arxiv.org/abs/1408.5093

[5] Nidhi Kalra and Susan M Paddock. 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94 (2016), 182–193.

[6] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. 2017. Playing for benchmarks. In *International Conference on Computer Vision (ICCV).*

[7] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision.* Springer, 102–118.

[8] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *CVPR.*

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[10] Alireza Shafaei, James J. Little, and Mark Schmidt. 2016. Play and Learn: Using Video Games to Train Computer Vision Models. *CoRR* abs/1608.01745 (2016). arXiv:1608.01745 http://arxiv.org/abs/1608.01745

[11] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

[12] Yang Zhang, Philip David, and Boqing Gong. 2017. Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. *CoRR* abs/1707.09465 (2017). arXiv:1707.09465 http://arxiv.org/abs/1707.09465