Understanding the Influence of Rendering Parameters in Synthetic Datasets for Neural Semantic Segmentation Tasks

Manuel Silva, Omar A. Mures, Antonio Seoane, and Jose A. Iglesias-Guitian

CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Civil Engineering Department, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: j.iglesias.guitian@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.47*

Abstract: Deep neural networks are well known for demanding large amounts of training data, motivating the appearance of multiple synthetic datasets covering multiple domains. However, synthetic datasets have not yet outperformed real data for autonomous driving applications, particularly for semantic segmentation tasks. Thus, a deeper comprehension about how the parameters involved in synthetic data generation could help in creating better synthetic datasets. This work provides a summary review of prior research covering how image noise, camera noise and rendering photorealism could affect learning tasks. Furthermore, we presents novel experiments aimed at advancing our understanding around generating synthetic data for autonomous driving neural networks aimed at semantic segmentation.

1 Introduction

In recent years, deep neural networks have become the prevailing solution for addressing most of the computer vision challenges, like object detection, image classification, and semantic segmentation (Janai et al., 2020). However, the progress of deep neural networks is often hampered by the scarcity and quality of available data. This becomes even more drastic in the case of semantic segmentation, where per-pixel annotation of real-world data requires a large human annotation effort. To address these challenges, current research has focused on two orthogonal but complementary topics: data augmentation and techniques to reduce manual labeling efforts.

Specifically, semantic segmentation produces pixel-wise classification maps where each pixel is assigned a class label (e.g., car, sidewalk, road). Real-world datasets with pixel-wise labels are very expensive to annotate, primarily due to the human effort involved, in addition, human annotations can easily become inconsistent or less accurate after a certain period of time. Consequently, the community is increasingly leveraging the use of computer graphics and simulation to generate synthetic datasets as a practical alternative to simulate acquisitions in the real world, i.e. digital twins or digital reality (see Fig. 1). Synthetic datasets can, not only expedite the annotation process but provide potentially more precise and consistent training data.

A notable challenge in utilizing synthetic data is the presence of a performance gap between models trained exclusively on synthetic data and those trained on real-world data, called domain adaptation (Wilson and Cook, 2020). In this study, we aim to investigate the behavior of synthetic images generated for training semantic segmentation models in the autonomous driving domain (Iglesias-Guitian et al., 2019).

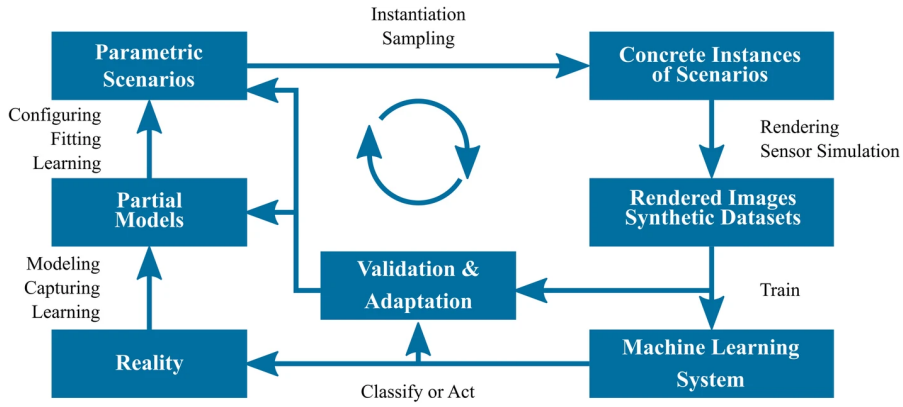

Figure 1: Digital reality approach (Dahmen et al., 2019) for supervised learning from synthetic data.

2 Related Work

Numerous studies have examined the relationship between image quality and the performance of neural network models. The great majority of studies are concerned with how artifacts in real images may affect the training of models for a specific task (Dodge and Karam, 2016; Kamann and Rother, 2020). In the context of real images, Pepik et al. (2015) showed that deep networks are not invariant to certain appearance changes, and demonstrated that the use of rendered data can serve to augment real image datasets. (Kamann and Rother, 2020) addressed the issue of image distortion impact over the semantic segmentation task, incorporating the autonomous driving dataset, Cityscapes (Cordts et al., 2016), alongside other datasets from diverse domains. After subjecting the data to various image distortions and noise, their study concluded that distortion artifacts preserving the fundamental characteristics of textures, such as various types of blur or brightness alterations, have a significantly lower impact than noise types that corrupt textures, like weathering effects or lossy compression. For a deeper and more general discussion about aspects affecting deep model's performance, we recommend the survey by Gawlikowski et al. (2023).

Studies on synthetic images . From the earliest applications of synthetic data for machine learning in autonomous driving (Pomerleau, 1991) to the widespread adoption seen today, synthetic datasets have emerged as a valuable alternative for data augmentation Paulin and Ivasic-Kos (2023).

One of the recurrent questions is how photorealism is really needed to achieve competitive model performance. For example (Movshovitz-Attias et al., 2016) addressed the question of how useful the photo-realistic rendering for synthetic data was for learning the viewpoint estimation task. The authors prepared three versions of a moderately sized synthetic dataset around cars using different levels of realism in terms of materials and lighting. The authors find that synthetic data could help achieve performances close to those of real data and that when mixing the two approaches the result could surpass that of using just real-world data. In addition, they also confirmed that increasing the training set size can help improve performance, but only up to a certain limit, probably due to the lack of variability in the source scenarios. Exploring the advantages of high-quality synthetic datasets, a study by McCormac et al. (2017) concluded that large-scale, high-quality synthetic datasets with task-specific labels could be more advantageous for pre-training models than generic real-world pre-training sources like ImageNet. This conclusion aligns well with previous findings (Tsirikoglou et al., 2017; Zhang et al., 2017) advocating the notion that pre-training models on physically based

rendered datasets with realistic lighting could substantially boost the performance for scene understanding tasks.

Previous works focused on how the photorealism of synthetic data could affect the neural networks, but they did not explore how other parameters used during the generation of the synthetic images could influence neural network's performance. In this sense, Liu et al. (2020) studied how different camera parameters generalize by using either real or synthetic datasets for object detection tasks, particularly cars. Their experiments evaluated pixel size, exposure control, color filters, bit depth, and post-acquisition processing. In the following, we summarized some of the preliminary conclusions drawn from their study. Notably, object size and distance within images showed limited correlation with generalization. Simulating diverse sensor pixel sizes was found to enhance generalization in specific contexts, suggesting its importance. Surprisingly, color filter generalization appeared symmetric, indicating potential cross-sensor applicability. Networks trained at distinct bit-depths displayed limited generalization, emphasizing bit-depth's significance. Exposure control algorithms played an important role, with good generalization observed, primarily affected by global exposure data quality. Finally, network performance exhibited robustness within certain gamma value ranges but deteriorated with significant deviations (e.g., > 0.1). Such preliminary conclusions shed light on potential factors influencing the generalization of the simulated camera parameters.

While previous research has been focused on what impact the synthetic image quality has in object detection tasks, a notable gap still remains in understanding how synthetic data could affect neural network performance in semantic segmentation for autonomous driving (Gómez et al., 2023; Khan et al., 2019).

3 Experiments

In order to address the aforementioned gap, we propose in this work a preliminary set of experiments focused on analyzing three key aspects while generating synthetic data for semantic segmentation: i) the influence of the training set size or number of images; ii) the camera sensor pixel size or image resolution; and iii) the number of samples per pixel(spp) used for path-traced physically-based rendering approaches. Our study aims to provide initial insights that can guide or inspire further research.

Figure 2: Various levels of Monte Carlo noise introduced by path tracing rendering. We compare how different samples per pixel could impact the semantic segmentation task.

To conduct our experiments, we adopt the widely recognized 19-class standard proposed in the Cityscapes dataset Cordts et al. (2016). We chose to have three different sources as synthetic

datasets: GTAV (Richter et al., 2016), Synscapes (Wrenninge and Unger, 2018) and a custom-tailored dataset for which we could control the samples per pixel for rendering. These three datasets will serve as sources, and two real-world datasets will serve as targets: Cityscapes and Mapillary.

Training set size. For this experiment our hypothesis is that, depending on the content and variation included in a synthetic dataset, the number of images required in the training set to achieve the best results for semantic segmentation may not follow the prevailing paradigm of *"more data leads to better performance"*. This phenomenon is particularly relevant when considering carefully crafted synthetic datasets, which may exhibit recurrent patterns or tend to overfit to specific layouts or scenarios. Our experiments tested the progressive reduction of the training set size for two source datasets, Synscapes and GTAV, while validating each of them on two different targets, Mapillary and Cityscapes. Please, check Fig. 3 and Sec. 4 for a detailed discussion.

Camera sensor size. In this experiment our hypothesis is that, depending on the simulated sensor resolution, certain details might be better represented at higher resolutions, while lower resolutions may better capture the general shape of larger objects without getting distracted by smaller details. Depending on the class of objects, different resolutions might obtain different results. For our experiments we tested the original (100%), half (50%), quarter (25%) and eighth (12, 5%) of the image resolution for both GTAV and Synscapes using Cityscapes as target (see Fig. 4).

Samples-per-pixel for path-tracing. With this experiment, we aim to test the performance of a neural network related to the path tracer's noise and the denoiser's smoothing effect. This type of noise is due to the variance of the Monte Carlo (MC) integration method used in path-tracer rendering engines, and it is not present in real images. Additionally, we aim to explore the influence of an AI denoiser, trained to reduce MC variance noise by adaptively smoothing the image and finding a trade-off to also preserve high-frequency details. For these experiments, we use a custom-developed dataset for autonomous driving based on an unbiased GPU path tracer, that offer us the possibility to control the aforementioned parameters. We use 425 images from this custom dataset, and we save renders at 8, 16, 64, 128, 256, and 512 spp, for both denoised and the original MC path traced samples. For a thorough discussion about these results, please see Fig. 4 and Sec. 4.

The framework utilized for all these experiments uses Detectron2's (Yuxin Wu et al., 2019) DeepLabV3+ Chen et al. (2018) semantic segmentation model and is based on the framework presented in (Gómez et al., 2023). We adopted the synth-to-real LAB space alignment defined in that framework to help reduce the synth-to-real domain gap.

4 Results

Training set size. In Figure 3 we illustrate the results about the training set size for both Synscapes and GTAV with two different targets. Basically, while both datasets caused great variance for lower number of images, the trend becomes more stable after 4k images with slight variations. These results seem to indicate that achieving equal or greater performance with a portion of the original dataset is possible.

Camera sensor size. In Figure 4 (left) we show the training set image resolution results for both Synscapes and GTAV using the Cityscapes validation set. Synscapes seems to find a trade-off while varyinf its image resolution, except for cases below one eighth of its original size. On

the contrary, GTAV presents a sort of sweet spot around using half the original resolution. Given the nature of GTAV, this result could support our notion that lower resolutions may help larger objects more visible in the image, resulting in a slightly better result.

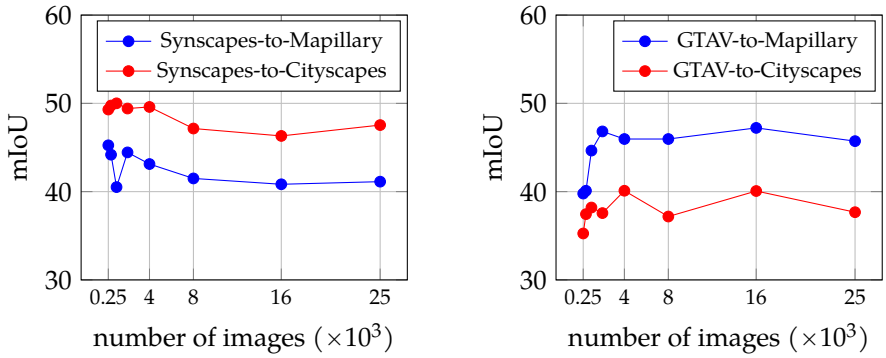

Figure 3: Semantic segmentation (mIoU) w.r.t. the number of images in Synscapes(left), and GTAV(right).

Samples-per-pixel for path-tracing. In Figure 4 (right) we show the performance while varying the number of samples per pixel when generating our custom urban dataset for both the original MC path-traced result and their respective denoised images. The differences between each run are very small and could be explained because of the standard deviation of the experiments. It's worth noting that the MC noise version appears to perform better with fewer samples, while the denoised version appears to function better with more samples. This could be explained by the network preferring the MC noise over the smoothed information supplied by the denoiser at lesser resolution. The denoiser works better at higher resolutions and does not remove as much texture detail as it does at lower levels, which may account for the chart's optimum performance at 256 spp.

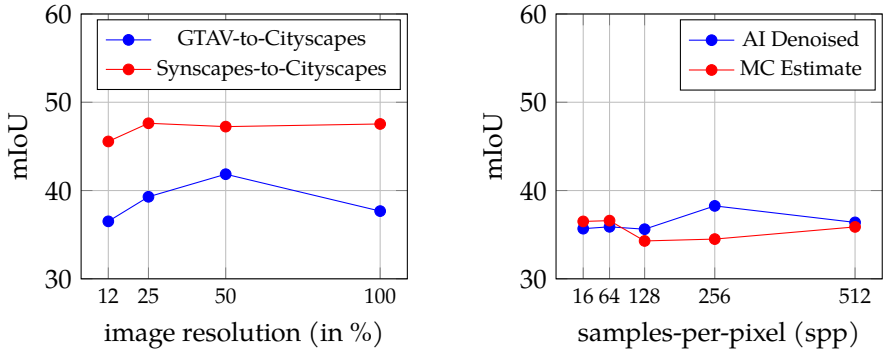

Figure 4: Semantic segmentation (mIoU) w.r.t. image resolution(left) and samples-per-pixel (right). The experiment on spp, used our custom generation process that allow us to control the spp. The two experiments shown here perform their validation using Cityscapes as target.

5 Conclusions

In this study, we reviewed the literature covering the impact that image quality has on neural network performance. We focus on works covering synthetic datasets and performed a preliminary set of experiments to gain new insights around the semantic segmentation task. Our initial findings suggest that similar results for different training set sizes can be obtained, probably due to the lack of variation in common autonomous driving synthetic datasets. Also, different image resolutions may have a positive impact on the performance. And finally, when using Monte Carlo path-tracing, it may be better to use denoising at higher sample rates, while MC noise might be preferable than using a denoiser at lower sampling rates for segmentation networks.

As future work we would like to confirm the initial findings of this work by using more rendering parameters and more experiments studying how the synthetic generation process may boost neural networks' performance.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation (AEI/PID2020-115734RB-C22). We also want to acknowledge Side Effects Software Inc. for their support to this work. J.A. Iglesias-Guitian also acknowledges the UDC-Inditex InTalent programme, the Ministry of Science and Innovation (AEI/RYC2018-025385-I) and Xunta de Galicia (ED431F 2021/11). CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Bibliography

L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(04):834–848, apr 2018.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

T. Dahmen, P. Trampert, F. Boughorbel, J. Sprenger, M. Klusch, K. Fischer, C. Kübel, and P. Slusallek. Digital reality: a model-based approach to supervised learning from synthetic data. *AI Perspectives*, 1(1):1–12, 2019.

S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016.

J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pages 1–77, 2023.

J. L. Gómez, G. Villalonga, and A. M. López. Co-training for unsupervised domain adaptation of semantic segmentation models. *Sensors*, 23(2):621, Jan 2023.

J. Iglesias-Guitian, G. Ros, V. Kokkevis, J. Alvarez, Y. Lee, and P. Slusallek. Computer graphics for autonomous vehicles. In *ACM SIGGRAPH Frontiers Workshop*. ACM SIGGRAPH, 2019.

J. Janai, F. Güney, A. Behl, and A. Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3): 1–308, 2020.

C. Kamann and C. Rother. Benchmarking the robustness of semantic segmentation models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8825–8835, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

S. Khan, B. Phan, R. Salay, and K. Czarnecki. Procsy: Procedural synthetic dataset generation towards influence factor studies of semantic segmentation networks. In *CVPR workshops*, volume 3, page 4, 2019.

Z. Liu, T. Lian, J. Farrell, and B. A. Wandell. Neural network generalization: The impact of camera parameters. *IEEE Access*, 8:10443–10454, 2020.

J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017.

Y. Movshovitz-Attias, T. Kanade, and Y. Sheikh. How useful is photo-realistic rendering for visual learning? *ArXiv*, abs/1603.08152, 2016. URL *https://api.semanticscholar.org/CorpusID:12684994*.

G. Paulin and M. Ivasic-Kos. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial Intelligence Review*, pages 1–45, 2023.

B. Pepik, R. Benenson, T. Ritschel, and B. Schiele. What is holding back convnets for detection? In *Pattern Recognition: 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings 37*, pages 517–528. Springer, 2015.

D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.

S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 102–118, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46475-6.

A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger. Procedural modeling and physically based rendering for synthetic data generation in automotive applications. *arXiv preprint arXiv:1710.06270*, 2017.

G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

M. Wrenninge and J. Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *CoRR*, abs/1810.08705, 2018. URL *http://arxiv.org/abs/1810.08705*.

A. K. Yuxin Wu, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2, 2019. URL *https://github.com/github-linguist/linguist*.

Y. Zhang, S. Song, E. Yumer, M. Savva, J. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5057–5065, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.